BIRLA CENTRAL LIBRARY
PILANI [RAJASTHAN]
Class No 621.384
Book No. J63F
Accession No 30525

PHYSICS

Its Laws, Ideas, and Methods

ALEXANDER KOLIN

Assistant Professor of Physics in the College
and in the Division of Biological Sciences
The University of Chicago

FIRST EDITION
THIRD IMPRESSION

McGRAW-HILL BOOK COMPANY, INC.
NEW YORK TORONTO LONDON · 1950

PHYSICS

Copyright, 1950, by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

THE MAPLE PRESS COMPANY, YORK, PA.

To My Students

CONTENTS

Preface . . . vii
Symbols Used in This Text . . . xi
List of Commonly Used Greek Letters . . . xiii
Foreword for the Student . . . xv

PART A. PHENOMENOLOGICAL APPROACH IN PHYSICS

MATTER IN MOTION AND ENERGY

1. Introduction . . . 3

I *Mechanics of Points and of Rigid Bodies*

2. Kinematics: 1. Linear Motion . . . 12
3. Kinematics: 2. Angular Motion and Curvilinear Trajectories . . . 25
4. Dynamics: 1. Newton's Laws (First and Second) . . . 33
5. Dynamics: 2. Newton's Third Law and Conservation of Momentum . . . 48
6. Dynamics: 3. Similarity between Effects of Gravity and of Inertia . . . 62
7. Dynamics: 4. Vectors and the Generalization of the Concept of Acceleration 72
8. Dynamics: 5. Rotational Motion . . . 89
9. Work and Energy: 1. . . . 105
10. Work and Energy: 2. . . . 119

II *Mechanics of Fluids*

11. Fluids at Rest . . . 132
12. Fluids in Motion . . . 151

III *Vector Fields*

13. Fields of Force: 1. The Gravitational Field . . . 175
4. Fields of Force: 2. Electrostatic and Magnetostatic Fields . . . 184
15. Fields of Force: 3. Behavior of Matter in Electric and Magnetic Fields . . 206

IV *Heat as a Form of Energy*

16. The Concepts of Temperature and of Quantity of Heat . . . 223
17. Transformations of Phase and Transfer of Heat . . . 240
18. Thermal Behavior of Gases . . . 257

PART B. THEORETICAL APPROACH IN PHYSICS USING MECHANICAL MODEL THEORIES

THE ATOMIC THEORY OF ELECTRICITY AND MATTER AS AN EXAMPLE OF A MECHANICAL EXPLANATION OF PHYSICAL PHENOMENA

I *As a Mechanical Explanation of Heat Phenomena*

19. Kinetic Molecular Theory: 1. Qualitative Part . . . 275
20. Kinetic Molecular Theory: 2. Quantitative Part . . . 290

II *As a Mechanical Explanation of Electrical Phenomena*

21. Electricity in Motion: 1. Migration of Ions in Liquids . . . 309
22. Electricity in Motion: 2. Atomicity of Electricity, Electronic Conduction, Electric Circuit Laws . . . 336

III *As a Mechanical Explanation of Electromagnetic Phenomena*

23. Interaction between Magnetic Fields and Moving Electric Charges 359
24. Interaction between Magnetic Fields and Moving Uncharged Bodies. . . . 392

PART C. TRANSITION FROM A MECHANICAL MODEL THEORY TO AN ELECTROMAGNETIC FIELD THEORY

THE WAVE THEORIES OF SOUND AND LIGHT

I *Periodic Motion of Matter and Wave Propagation in Material Media*

25. Oscillatory Motion. 427
26. Wave Propagation. 441
27. Sound as an Example of Wave Motion. 463

II *Wave Propagation in the Hypothetical Ether*

28. The Nature of Light. 495
29. Light as a Wave Motion 514
30. Formation of Optical Images 539
31. The Study of Spectra 586

III *Periodic Motion of Electric Charges and Propagation of Electromagnetic Waves*

32. Oscillations of Electric Charges 608
33. Electromagnetic Oscillations and Waves. Maxwell's Theory of Light . . . 635
34. Extension of the Electromagnetic Spectrum from Hertzian Waves to X Rays. 668

PART D. THE FAILURE OF THE CLASSICAL MECHANICAL AND ELECTROMAGNETIC THEORIES AND THE REVISION OF THE FUNDAMENTAL ASSUMPTIONS OF PHYSICS

I *Inability of Classical Electrodynamics to Explain Emission and Absorption of Light*

35. Evolution of the Theories of Light and Matter 697
36. Bohr's Atomic Theory 721

II *Replacement of Deterministic Description by Statistical Laws*

37. Wave Mechanics 742

III *Modification of Our Ideas on the Immutability of Elements and on the Permanence of Mass and Energy*

38. Spontaneous Transmutation of Elements. 768
39. Artificial Transformation, Creation, and Annihilation of the Building Blocks of Matter. 797

IV *Revision of Our Concepts of Space and Time*

40. The Theory of Relativity. 836

Tables. 867

References for Further Study. 873

Index 875

PREFACE

This book is a thoroughly revised edition of "Introduction to Physics" (University of Chicago Press, 1948) which was written by the author for the college course in physics at the University of Chicago. This physics course has been designed to accomplish a double purpose:

1. To satisfy the requirements of the Division of Biological Sciences in regard to the training in physics of students entering that division.
2. To satisfy the general education objectives of the college in regard to the training of students in the methods of the physical sciences.

These two objectives can be reconciled without conflict, since the intellectual training provided by the study of the scientific method of inquiry is of equal benefit to students of all sciences and highly desirable for students of liberal arts. The relatively numerous examples of biological applications of physics are intended to enhance the interest of the biology students, but they may prove no less stimulating than examples from technology, astronomy, or geology to nonbiologists. (No attempt has been made to present a complete survey of applications of physics to biology.)

The emphasis of the book is on the elucidation of the scientific method as exemplified in its application to physics, and on the understanding of physics as a logically coherent science, whose objective it is to establish the basic laws of nature which form the foundation of all natural sciences. The transferable features of the methods used in physics are emphasized so as to stimulate the student to apply them to other fields of study. The relation of physics to other natural sciences and its role in our civilization are illustrated in a number of examples.

Exercises in scientific thinking are provided in numerous questions and problems which fall into the following categories:

1. Questions which test whether or not the student has understood what he has read.
2. Questions and problems training in analytical thinking involving an analysis of the concepts and theories of physics.
3. Questions developing a skill in deductive logical thinking by solution of quantitative problems.
4. Questions training the student in synthetic thinking, such as those which call upon his ingenuity by requiring him to design experimental methods.
5. Problems which challenge the student to apply physics to biological questions.

The organization of subject matter in this text differs from the conventional organization of physics. The subjects are arranged according to methods of treatment and the nature of ideas. The student begins with the simplest and most concrete approach and progresses gradually to more complex and more abstract ideas and methods. Part A deals with the straightforward mathematical formulation of physical laws inferred by induction; the use of hypotheses is almost completely avoided. The student thus becomes familiar with several branches of physics which seem to have no logical connection with each other. In Part B the theoretical method is introduced. The introduction of the atomic theory of matter and electricity enables us to establish a logical connection between the laws of mechanics, heat, and certain electrical manifestations. This success suggests the aim of explaining all physical phenomena on a mechanical basis. In Part C this aim is given up as unattainable. After an initial study of oscillations of matter and electricity which are amenable to a common mathematical treatment, a mechanical wave theory of light is developed but finally dropped as untenable and replaced by an electromagnetic wave theory. In Part D the inadequacy of the concrete visualizable theories of classical physics in the realm of high speeds and small dimensions compels us to develop new theories which turn out to be quite abstract, bizarre in their consequences, and eminently successful in predicting new laws.

There is no introductory chapter designed to summarize the philosophy of science. Methodological considerations are deliberately spread over the entire text and usually introduced at a point where concrete examples can be used to illustrate the general statements. The first introductory chapter is merely a "first installment" of a sequence of discussions concerning our sources of knowledge and methods of inference.

In a less intensive course the numerous passages set in small type may be omitted without loss of continuity. This applies mostly to the appendices and the biological and technological examples, as well as to the last chapter on The Theory of Relativity.

I wish to thank Dr. R. Sternheimer and Professor L. Nedelski for reading critically parts of the manuscript of the previous edition. Thanks are also due to G. Backus, H. R. Paneth, and A. H. Rosenfeld for valuable criticisms and suggestions; to Professor R. W. Gerard for his kindness in reading and commenting on the section on electrophysiology; and to E. H. Denton for his kind assistance in the preparation of the index. I also wish to thank the many authors and firms who have permitted me to reproduce their original photographs and drawings, acknowledgment of which is to be found in the legends.

ALEXANDER KOLIN

CHICAGO, ILL.
September, 1950

COMMONLY USED GREEK LETTERS

α	alpha	θ	theta	ρ	rho
β	beta	κ	kappa	Σ, σ	sigma
γ	gamma	λ	lambda	τ	tau
Δ, δ	delta	μ	mu	ϕ	phi
ϵ	epsilon	ν	nu	ψ	psi
η	eta	π	pi	Ω, ω	omega

FOREWORD FOR THE STUDENT

The purpose of this book is not merely to describe a number of ingenious "gadgets" and rules of computation but rather to provide you with examples of and exercises in scientific thinking and to acquaint you with the basic laws of nature. The so-called "method of scientific inference" used in the development of the science of physics is not limited to physics alone but is useful in other sciences as well, and many of its practices could be profitably transferred to human affairs. The main purpose of this book is to illustrate how logical thinking and creative imagination lead from the known to the prediction and comprehension of the unknown. But it is to be borne in mind that one cannot acquire mastery in scientific thinking by merely reading a book, any more than one can become proficient in public speaking by listening to speeches. An essential part of your efforts should be devoted to the solution of questions and problems and to the formulation of questions of your own.

The nature of the difficulties which you will encounter varies throughout the book. At the beginning your efforts will be taxed by the unfamiliar method of mathematical reasoning. But the mathematics remains on a uniformly elementary level throughout the text and will soon cease to be a difficulty. As you progress, you will find the difficulty to lie more and more in the use of abstract concepts and of indirect arguments.

As you read and try to understand the text, it is well to bear in mind what is meant by understanding. To understand a *concept* you must know its operational definition (see Chapters 1 and 2). To understand a *definition* you must be familiar with the concepts in terms of which the definition is formulated. Understanding of a *phenomenon* consists in showing the phenomenon to be a special manifestation of a general law. We can say that we understand a *law* if we can deduce it from a more general law or from a theory (which thus plays the role of a more general law). And, finally, the understanding of a *theory* consists in recognizing which are its basic assumptions, the empirical basis for these assumptions, and the logical consequences deduced from these assumptions.

The following procedure is recommended for your studies:

1. Before reading the chapter glance at the summarizing questions at the beginning of the chapter. This will acquaint you with the salient points of the chapter.
2. Read the chapter rapidly to get the main ideas.
3. Read the limited assignment carefully.
4. Write down answers to all summarizing questions. This tabulation will comprise a summary of the main ideas of the chapter. This is not an

exercise in thinking, since most of the answers to these questions are to be found in the text.

5. Make an abstract from the reading assignment, tabulating all (*a*) definitions, (*b*) laws, (*c*) important (framed) equations, and (*d*) units which occur in the assignment. This task is facilitated by the fact that, as a rule, laws and important definitions are printed in boldface type. This summary will be very helpful in reviewing the chapter and as a quick reference in solving problems.

6. Answer a selection of the Questions and Problems.

Some of the questions are deliberately meaningless. The student should be alert to recognize such questions and to subject them to criticism. The starred questions are intended for the most ambitious of the students.

At the end of some chapters are given references on biological applications of physics which are discussed in the chapter. A list of references on various branches of physics which the student may consult for specialized information is given at the end of the book.

> . . . I frame no hypotheses; for whatever is not deduced from the phenomena is to be called an hypothesis; an hypothesis, whether metaphysical or mechanical, has no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterwards rendered general by induction . . . .
>
> *Newton*

PART A: Phenomenological Approach in Physics

In this part of the book we confine ourselves to the observation of phenomena, the definition of concepts in terms of which they are described, and the formulation of generalizations (laws) which are inferred from a limited number of observations. We become acquainted with the mathematical method of formulating physical laws as functional relations between variable quantities. It is then shown how answers to specific questions can be deduced from general laws.

The procedure outlined above is illustrated in the example of some phenomena which are accessible to direct observation, such as motion, deformation, and heating of bodies. We begin by establishing the basic concepts and methods for the description of the motion of bodies (Chaps. 2 and 3) and proceed then to study how interaction between bodies affects their motion (Chaps. 4 to 8).

In Chaps. 9 and 10 we infer a new general law (the law of conservation of energy) whose validity is not limited to the motion of bodies but rather extends to all processes occurring in nature. This law thus appears not only as a conceptual bridge between the laws of mechanics and those of other branches of physics but as a bridge between physics and other natural sciences.

In Chaps. 11 and 12 the laws of mechanics, established in previous chapters for point masses and rigid bodies, are applied to deformable, continuous matter (fluids). In the study of fluid motion we develop the concept of a "velocity field," which is generalized to the picture of a "vector field," and methods are found to represent such "fields" graphically. These ideas are used in the study of electric and magnetic interactions of bodies at rest (Chaps. 13 to 15).

In Chaps. 16 and 17 heat phenomena which bear no recognizable relation to mechanics are discussed. And in Chap. 18 the mechanical and thermal properties of compressible fluids are studied. The interconversion between heat and mechanical energy suggests the possibility of linking logically the seemingly unrelated fields of heat and mechanics. The laws governing the thermal behavior of gases give clues as to how such a relation may be established. The main clue is the dependence of certain thermal properties of a gas on its chemical molecular weight. This clue is followed up in Part B of the book leading to a new approach (the theoretical approach) in the exploration of the laws of nature.

> Orderly arrangement is the task of the scientist. A science is built out of facts just as a house is built out of bricks. But a mere collection of facts cannot be called a science any more than a pile of bricks can be called a house.
>
> *Poincaré*

CHAPTER 1

INTRODUCTION

What Is a Science?

If you ask a physicist how fast a falling blue glass marble, released from rest, will move 1 sec after its release, he will give you the answer without looking it up in a table or without performing a measurement. How does he know it? He knows it from the generalization that (neglecting air resistance) all objects (blue marbles, green marbles, as well as grand pianos) fall at a specific locality with the same acceleration (known to him from experiments) and from the further generalization that the velocity after falling 1 sec is numerically equal to that acceleration.

If you ask him, "How much will I weigh if I stand on a tower whose height is equal to the earth's radius?" he will say, "You will weigh one-fourth of what you weigh at the earth's surface." How does he know it? Nobody has ever built such a tower, and nobody ever weighed objects at such a height. He knows it from the generalization that the weight of a body decreases in inverse proportion to the square of its distance from the earth's center.

If you ask a chemist what will happen to a new hormone which you have just synthesized at a temperature of 5000°C, he will tell you without hesitancy that it will decompose into its constituent elements, although he had never heard of this hormone before. Again, this answer is based on a generalization, namely, that all chemical compounds decompose before they reach a temperature of 5000°C.

We can infer from these examples that a science is a collection of generalizations. The aim of any science is to establish a set of generalized statements (laws) from which the answer to any particular question could be obtained.

This aim may not sound very worth while. There is a large store of profound generalizations, such as "At night all cats are gray" or "Still waters run deep," on which we seldom like to rely. As we shall see presently, however, the generalizations of science, unlike the aforementioned ones, are not inspired by the imagination of a poetic mind but are carefully established by methods which inspire in us confidence in their validity.

Divisions between the Sciences

The enormous amount of accumulated knowledge about nature has been classified into a number of specialized fields according to the nature of the phenomena under consideration. The divisions are usually not sharp and in many cases obviously arbitrary. In some instances a gradual transition may be traced. For example, the separation between biological and physical sciences does not appear so sharp when we think of the tobacco mosaic virus (which represents the lowest form of life), which can be crystallized like so many other substances the study of which forms the subject matter of chemistry.

The limit between chemistry and physics cannot always be drawn sharply either. One of the major concerns of contemporary physics, for example, is to produce a variety of new chemical elements by means which have been traditionally called physical methods.

Physics

The subject matter of physics is the study of laws according to which matter interacts with matter or energy. For instance, matter can attract matter due to gravity or due to electrification or magnetization. Such interactions may lead to changes in motion, which is the subject matter of **mechanics.** A hot body may heat a cooler one by conduction or by radiation. Such interactions are studied in the chapter on **heat.** We shall see, for instance, that tiny fragments of amber may under certain conditions repel other similar fragments of amber. Such interaction is called **electrical.** The electrical interaction is modified by the state of relative motion of the two interacting bodies. The study of this modification forms the subject matter of **electromagnetism.** The study of **sound** leads to the consideration of interaction between particles of matter engaged in periodic motion. And finally, the study of **optics** is concerned with the processes of emission, absorption, propagation, and distribution of light and of invisible radiations analogous to it.

Relation of Physics to Other Sciences

It is not difficult to see how physics is related to other natural sciences. All natural sciences deal with material bodies which interact with each other. As we have seen, the study of such interactions belongs in the domain of physics. This fact establishes the link between physics and the other natural sciences. Furthermore, all information on nature is obtained by the use of our senses. The subdivision of physics into branches like heat, sound, and light points clearly to the fact that physics deals with the agents and changes which affect our senses. Physics develops the means of sharpening our senses, as it were. When a chemist mixes reagents, he does not rely on his muscle sense to estimate their weight; he uses a balance. When he gets a dark blue solution in the course of a reaction, he can measure the "depth" of the color by electro-optical means. When a geologist is prospecting for iron ore, he employs magnetic methods developed by physicists. An astronomer detects heat radiation from distant stars by the use of electrical circuits and optical instruments. The biologist uses electrical amplifiers and oscilloscopes to detect electric variations in tissues which are too weak to reveal themselves to us through an electric shock. Similarly, our powers of perception have been extended beyond the limits accessible to our senses by the use of X rays, ultraviolet and infrared light, microscopes, telescopes, electron microscopes, and electronic sound amplifiers, which enable us to establish generalizations in all natural sciences, even in domains which would have remained invisible and inaudible otherwise.

But acquisition of knowledge about nature is not accomplished by perception alone. It is achieved by a combination of perception and reasoning. The contribution of physics was not limited to the sharpening of our senses but was equally important in perfecting a mental device, the scientific method of inference, which has been adopted by all natural sciences.

Methodological Contributions of Physics

It was no mere coincidence that we owe this development largely to a physicist (Galileo). In physics we are capable of varying and simplifying conditions of observation as in no other natural science. If a physicist is interested in the relation between the properties A and B of a body, he can frequently keep all other properties of his system invariant, focusing his attention on the relation between A and B.

What is this so-called "scientific method of inference"? In what way does it differ from the way we acquire knowledge in everyday life? We shall see presently that the scientific method is essentially what normal people use in daily life. They are not only careful to draw logical conclusions from given premises but they also examine the validity of their premises carefully. Nothing was wrong with the logic of the physical sciences before Galileo, but the premises for its conclusions were frequently derived from human imagination bearing no relation to nature. Galileo established the rule that *no statement has a place in science which* *is not derived from direct observation or is a logical conclusion from premises based on observation.*

Let us now examine in detail the process through which we obtain information about nature by experience.

The Method of Induction. Suppose we are carrying a heavy stone while wading across a shallow pond. We drop the stone accidentally and pull it out of the water. On this occasion we notice that the stone feels lighter under water. This single observation is not very important. Nobody is especially interested in the behavior of this particular stone in that particular pond. But we may be moved to drop another stone and again a third into the pond. We shall find that all of the stones we dropped feel lighter. The more stones we test the more firm becomes our belief that all stones lose weight when submerged in our pond. We can repeat this experiment in other bodies of water. We shall obtain the same results and will extend our generalization by saying that all stones submerged in water anywhere lose weight. But is this the most general statement we can arrive at as a result of such experiments? We can refine our means of estimating the weight of a body and try bodies other than stones and liquids other than water or even gases. We shall find in an extensive series of such trials that *all* bodies submerged in *any* liquid or gas weigh less than in empty space.

In making this assertion we obviously went far beyond the range of our experience, since we tested only a few particular bodies in a few particular fluids.

What we just illustrated was the process of generalization or **induction** through which we infer a general law from a limited number of observations. All knowledge of nature in science as in everyday life is gained by this process.

Evidently, we can never be certain about an induction, since it always makes a statement going beyond our range of experience. Herein lies its value and at the same time its weakness.

But if we cannot say that an induction is certain, how can we express our degree of confidence in it? We can say that it is highly probable. Our belief in induction grows with the number of confirmations which support it; this makes the validity of the induction more probable.

As was pointed out by Poincaré, if an astronomer predicts an eclipse of the moon on a certain date about 2 years from now, he makes the tacit assumption that the moon, in the intervening period, will be subject to the same influences that govern it now. It is not impossible that a comet may invade our celestial neighborhood and disturb the course of the moon, overthrowing the prediction of the astronomer, but it is highly improbable. If we had no confidence in the improbability of such events, we could venture no scientific predictions and generalizations and there would be no science.

We are guided by similar subconscious probability considerations when we interpolate or draw a smooth curve through a sequence of determined experimental points. Suppose we draw a fever curve of a patient, taking the temperature every 15 min, and find the following series of data: 96.9°F, 97.3°F, 97.7°F, 98.1°F, 98.4°F, 98.7°F. We conclude that his temperature is rising at a fairly uniform rate and connect these points by a smooth ascending line. But what right have we to do this? Perhaps his temperature jumped temporarily to 97.6°F between the first two readings and dropped to 97.2°F between the third and fourth readings for 2 or 3 min? This may not be impossible, but we dismiss it as quite improbable.

None of our knowledge about nature can be considered as absolutely certain but at best as highly probable.

Inference by Deduction. For what purpose does the scientist amass a stockpile of generalizations? His set of generalizations, or **laws of nature,** as we shall call them, enables him to infer the answer to particular questions by a process of reasoning which is known as inference by **deduction.**

We shall give an example of a deduction in the classic form of a **syllogism** (which, by the way, is not an all-embracing pattern of deduction):

General proposition (major premise): All strong acids dissolve zinc.
Particular proposition (minor premise): HCl is a strong acid.
Deduced proposition (conclusion): Hence, HCl dissolves zinc.

The solution of any particular problem in physics in which the answer to a special question is inferred logically from general laws is a deduction. The solution of physics problems will provide ample exercise in deductive reasoning.

While experience is the basis of fundamental laws inferred by induction, the experimental method is not the only one by which general laws of nature can be established. From two or more propositions which spring from induction, new general propositions, new laws, can be inferred by pure logical reasoning. The new generalizations, if valid, should lead to deductions verifiable by experiment. The experimental physicist furnishes thus the facts on which inductions are based, and he tests the validity of the new laws which the theoretical physicist may infer from such inductions.

Concepts. Let us now examine a typical general law of physics, Newton's law of gravitation: **Two bodies attract each other with a force**

which is proportional to the mass of each body and is inversely proportional to the square of the distance between their mass centers.

This law (like any other law) is expressed in terms of certain concepts which have to be defined to make the law intelligible to everybody. There are concepts such as force, mass, distance, mass center, some of which are used in daily life with varying connotations. Their meanings must be fixed by definitions to avoid ambiguity in the interpretation of the meaning of the law.

But how do we define a term? It seems we can define it only in terms of other terms. And don't we have to define those "other" terms to avoid ambiguity? But then what would stop us from continuing definition of terms ad infinitum?

It appears reasonable to stop at some concepts the meanings of which appear intuitively evident and to adopt them as basic, undefined concepts in terms of which other concepts are to be defined. Such concepts are, for instance, time, length, and mass.

This procedure may look at first unsatisfactory. If we don't define the meaning of these concepts, how can we be sure that all physicists will mean the same thing when they speak, say, of time?

Operational Definitions. This difficulty can be avoided by a procedure which is called **operational definition,** which describes how the physical entity associated with the concept in question could be determined (or measured) instead of fixing its meaning in terms of other concepts. For instance, though we cannot clarify the meaning of time in terms of simpler, more evident concepts, we can prescribe a simple operation through which a time interval can be measured by means of a clock, or we can prescribe how a length is to be measured with a yardstick without stating what length is.

According to the school of thought represented by Bridgman, operational definitions should be given for all useful physical concepts, and concepts for the determination of which we can prescribe no feasible operation should be given up as having no physical reality, being undetectable by physical means. This view is not accepted without reservation by all scientists. An alternative view is expressed in the following quotation by Teller:

> I believe that the concept of operational definition is slightly overstressed and that the role of intuition is thereby somewhat underestimated. While it is true that we are striving to eliminate unsupported intuition in those parts of physics which are now fully understood, it is my conviction that intuition cannot be eliminated in the growing part of any science.

The importance and usefulness of an operational analysis of a scientific statement are due to the fact that it enables us to discover occasionally hidden, misleading hypotheses associated with basic concepts. Such was, for instance, the case with Newton's notion of absolute time. Newton writes in his "Principia": "Absolute time and mathematical time of itself, and from its own nature flows equably without regard to anything external, and by another name is called duration." Now, there is no physical operation which would enable us to detect that there is such a thing in nature as "absolute time," that it "flows from its own nature," and that it "flows equably." Hence, one has abandoned the concept of absolute time in favor of an operationally defined clock time.

The initiation of operational analysis of physical concepts was one of the great contributions of Einstein. Operational analysis is not necessarily limited to physical concepts. It can be applied to concepts in any field, and Bridgman suggests that much meaningless talk and disagreement would be avoided if the practice of an operational analysis of concepts were extended beyond the field of physics.

As an exercise in operational analysis the student may try to decide which of the following statements and questions are operationally meaningless:

1. All space and matter are permeated with an undetectable ethereal substance.
2. Is the sensation which I experience when I see green the same as that which you experience when you see green?
3. It is impossible to prove logically the validity of the rules of logical reasoning.
4. The entire universe is expanding so that everything in the universe has all its linear dimensions doubled every $3\frac{1}{2}$ months.
5. Dr. X: "I am sure the old man recovered because I gave him a large dose of my new drug Anticuridine." Dr. Y: "This is nonsense; he would have recovered just as fast if you had not given him anything."
6. Patient: "Will this new drug enable me to survive my illness, Doctor?" Doctor: "I don't know, but it will about double your chances of survival."
7. If there were a species of flat animals inhabiting the surface of a sphere, unable to leave this surface, their habitat would be a two-dimensional one. Could they nevertheless infer from experience that space is three-dimensional?

APPENDIX

Newton's "Principia," Book III[†]

Rules of Reasoning in Philosophy:

Rule I

We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearance.

To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.

Rule II

Therefore to the same natural effects we must, as far as possible, assign the same causes.

As to respiration in a man and in a beast; the descent of stones in Europe and in America; the light of our culinary fire and of the sun; the reflection of light in the earth, and in the planets.

Rule III

The qualities of bodies, which admit neither intensification nor remission of degrees, and which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever.

For since the qualities of bodies are only known to us by experiments, we are to hold for universal all such as universally agree with experiments; and such as are not liable to diminution can never be quite taken away. We are certainly not to relinquish the evidence of experiments for the sake of dreams and vain fictions of our own devising; nor are we to recede from the analogy of Nature, which is wont to be simple, and always consonant to itself. We no other way know the extension of bodies than by our senses, nor do these reach it in all bodies: but because the hardness of the whole arises from the hardness of the parts we therefore justly infer the hardness of the undivided particles not only of the bodies we feel but of all others. That all bodies are impenetrable, we gather not from reason, but from sensation. The bodies which we handle we find impenetrable, and thence conclude impenetrability to be an universal property of all bodies whatsoever. That all bodies are movable, and endowed with certain powers (which we call the inertia) of persevering in their motion, or in their rest, we only infer from the like properties observed in the bodies which we have seen. The extension, hardness, impenetrability, mobility, and inertia of the whole, result from the extension, hardness, impenetrability, movability and inertia of the parts; and hence we conclude the least particles of all bodies to be also all extended, and hard and impenetrable, and movable, and endowed with their proper inertia. And this is the foundation of all philosophy. Moreover, that the divided but contiguous particles, that remain undivided, our minds are able to distinguish yet lesser parts, as is mathematically demonstrated. But whether the parts so distinguished, and not yet divided, may, by the powers of Nature, be actually divided and separated from one another we cannot certainly determine. Yet, had we the proof of but one experiment that any undivided particle, in breaking a hard and solid body, suffered a division, we might by virtue of this rule conclude that the undivided as well as the divided particles may be divided and actually separated to infinity.

[†] Reprinted from the Cajori edition of Newton's "Principia" by permission of the University of California Press.

Lastly, if it universally appears, by experiments and astronomical observations, that all bodies about the earth gravitate towards the earth, and that in proportion to the quantity of matter which they severally contain; that the moon likewise, according to the quantity of its matter, gravitates towards the earth; that, on the other hand, our sea gravitates towards the moon; and all the planets one towards another; and the comets in like manner towards the sun; we must, in consequence of this rule, universally allow that all bodies whatsoever are endowed with a principle of mutual gravitation. For the argument from the appearances concludes with more force for the universal gravitation of all bodies than for their impenetrability; of which, among those in the celestial regions, we have no experiments, nor any manner of observation. Not that I affirm gravity to be essential to bodies: by their *vis insita* I mean nothing but their inertia. This is immutable. Their gravity is diminished as they recede from the earth.

Rule IV

In experimental philosophy we are to look upon propositions inferred by general induction from phenomena as accurately or very nearly true, not-withstanding any contrary hypotheses that may be imagined, till such time as other phenomena occur, by which they may either be made more accurate, or liable to exceptions.

This rule we must follow, that the argument of induction may not be evaded by hypotheses.

CHAPTER 2

KINEMATICS: 1. LINEAR MOTION

a. How do we define motion; *i.e.*, what is the criterion whereby we can tell whether or not an object is moving?

b. If you were the only object in the universe, how could you detect whether or not you were moving?

c. If, besides yourself, there were another object in the universe seemingly approaching you, how could you distinguish whether (1) you are at rest while it moves toward you? (2) you are moving toward the object while it is at rest? or (3) both you and the object are in motion?

d. If you were in a closed, windowless train which is moving uniformly so smoothly as not to betray its motion, by what physical experiment could you decide whether or not you were in motion?

e. What is distance? How do you define it?

f. What is time? How do you define it?

g. What is velocity? How do you define it?

h. What is acceleration? How do you define it?

i. Can an object be accelerated without speeding up or slowing down?

j. If you are sitting in a fast-moving train, what meaning (if any) has the question: "Are you at rest or are you in motion?"

k. When do we call a quantity a vector? when a scalar?

l. In what sense are we justified in considering Galileo's work a turning point in the history of science?

m. In establishing a new law, do you use the process of induction or of deduction?

n. Are Galileo's formulas for freely falling bodies the same thing as Galileo's laws of free fall?

o. Can we ever predict the outcome of experiments (a priori) without any preliminary factual experience?

p. Is the assertion that heavier bodies fall faster inconsistent with either reasoning or experience?

q. Is the definition of instantaneous speed, $v = s/t$, valid only for uniformly accelerated linear motion?

r. Do nonuniformly accelerated motions occur and play an important role in nature?

Concept of Space

If our world were a motionless conglomeration of objects, we could not do any more in it than study the disposition of the perceived objects with respect to each other. Idealization of the observed forms would lead us to the development of geometry, the study of space. The concept of space is an intuitive concept referring to our ability to perceive objects apart from each other (at any given instant).

Concept of Time

But our world is not motionless. Most objects we look at are subject to change. They may change their position or their appearance. We are not dealing merely with fixed objects but with changes, with events. We find that our ability to order objects in space is not sufficient to cope with the study of changes. The consideration of events necessitates the establishment of an order of succession. Our ability to perceive events one after the other is the basis of our intuitive concept of time.

Operational Definitions

Our concepts of space and time do not lend themselves to a definition in terms of more familiar concepts. We have to accept them as undefined intuitive concepts. But we can define each of them "operationally"; *i.e.*, we can describe operations by which the relative position between two points in space can be ascertained in terms of a specified unit of comparison. And we can define an operation according to which the time interval between two events can be measured in terms of a standard time interval, or unit of time. Even though we *don't* know what time is, we do know what 10 sec are, namely, an interval 10 times as long as a standard time interval of 1 sec, and similarly 25 cm is a distance 25 times as long as the 1-cm standard.

Just as we measure a length by the operation of determining how many length standards we can place end to end alongside it, we can measure a time interval by counting the number of seconds which elapse between its beginning and its end.

Units of Length and Time

The unit of length is defined arbitrarily as the length of a preserved standard (*e.g.*, standard meter in Paris). The unit of time has been defined as a prescribed fraction of the mean solar day. The constancy of this time standard depends on the constancy of the rotation and revolution of the earth.

Kinematics

The study of the relative disposition of certain elements in space is the subject matter of *geometry*. The study of the changes in spatial disposition of geometric elements in the course of time is the subject matter of *kinematics*. Thus, kinematics can be considered as an extension of geometry, namely, a study of the disposition of geometrical elements in space and time, *i.e.*, more plainly, the study of motion.

Relativity of Motion

We can detect displacement or motion of a body only relative to another reference body. There is no absolute (nonrelative) way of detecting uniform linear motion. In any two laboratories which are moving uniformly in a straight line relative to each other, all laws of physics are the same, so that an occupant could not determine his state of motion by any experiment whatsoever within his box. He could become aware of his motion relative to the other laboratory only by looking out of a window.†

Undefined and Defined Concepts

As in geometry, so in kinematics, we start with the simplest concepts which cannot be defined in terms of still simpler ones but of which we have an intuitive conception. These are the concepts of length and time; we must content ourselves with their "operational definitions" (see pages 8, 13).

Velocity

In terms of these primary concepts we can now define the other fundamental concepts of kinematics: **velocity** and **acceleration.** Velocity describes the motion of a point, giving the direction of motion and specifying how fast it is.

Vectors. Velocity is the first representative of a class of physical magnitudes that we encounter which are not described completely unless we state their direction. Such quantities are called **vectors.**

Scalars. Their counterparts are the so-called **scalar** quantities, which have no directional meaning and are completely described by one datum, namely, their magnitude. Volume, time, temperature are scalars. Force, velocity, displacement are vectors.

Example. A kangaroo jumps 10 ft 35° north of east. These data define a displacement vector. Ten feet is the magnitude of the vector, and 35° north of east states its direction. In the case of velocity the magnitude of the vector has a special name: **speed.**

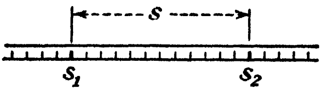


FIG. 2.1. Change of position of a point moving along a scale.

Linear Motion. The average velocity‡ in linear motion is defined as follows:

$$\bar{v} = \frac{\text{total displacement}}{\text{time interval}} = \frac{s}{t} = \boxed{\frac{s_2 - s_1}{t} = \bar{v}} \qquad (2.1)$$

$s_2 - s_1 = s$ represents the change of position of the point (Fig. 2.1). Thus, we can say: $\bar{v}$ represents the average change of position per unit time. (When we choose $t = 1$ sec, then $\bar{v} = s_2 - s_1$, that is, $\bar{v}$ is the change of position in a unit time interval.) Equation (2.1) is a definition equation. It is valid for nonuniform as well as uniform linear motion.

---

† The empirical and theoretical basis of this assertion will be discussed in detail in the chapter on The Theory of Relativity.

‡ The bar over a letter designates that it is an average value.

Accelerated Motion. Any motion in which the velocity varies in magnitude, direction, or both is called *accelerated motion.* Although in common usage acceleration implies a speeding up, in the physical sense of the word an accelerated body may also be slowing down (negative acceleration) or it may continue moving equally fast all the time, changing merely its direction of motion. For example, the motion of the tip of a watch hand is an accelerated motion.

Instantaneous Velocity. Is there any sense in speaking of an instantaneous velocity in the case of a nonuniform motion, where v is constantly changing in magnitude and possibly in direction? Let us limit ourselves at first to rectilinear motion. If we choose the points s_1 and s_2 close enough to each other, the body may take a very short time to move from s_1 to s_2. Then $\bar{v} = (s_2 - s_1)/t = (s_2 - s_1)/(t_2 - t_1)$, where $t_2 - t_1 = t$ is the short time interval which the body takes to move from s_1 to s_2.

It is customary to designate small differences by the Greek letter Δ: $s_2 - s_1 = \Delta s$ and $t_2 - t_1 = \Delta t$. Thus,

$$\boxed{v = \frac{\Delta s}{\Delta t}} \tag{2.2}$$

The bar has been omitted over v to indicate that it can be considered as the instantaneous velocity at the point s_1 if we choose s_2 indistinguishably close to s_1 by observing the motion for a short enough time interval Δt.

Thus, **the instantaneous velocity is defined as the average velocity over an infinitesimally short time interval.** (The case of nonlinear motion will be taken up later.) The unit of velocity is $\frac{\text{unit of distance}}{\text{unit of time}}$, for instance, centimeters per second, feet per second, feet per minute, miles per hour, etc.

Average Acceleration. The concept of instantaneous velocity is necessary for the treatment of accelerated motion, *i.e.*, motion in which the instantaneous velocity varies.

The definition of average acceleration in linear motion is formally analogous to the definition of average velocity:

$$\bar{v} = \frac{s_2 - s_1}{\Delta t} \tag{2.1}$$

Average acceleration:
$$\bar{a} = \frac{v_2 - v_1}{\Delta t} \tag{2.3}$$

Just as $\bar{v}$ = change of position per unit time,
$\bar{a}$ = change of velocity per unit time.

To get $\bar{v}$ we read the position s_1 at the beginning of the time interval Δt and the position s_2 at its end and divide $s_2 - s_1$ by the time interval.

We obtain $\bar{a}$ by the same operation except that, instead of the initial and final positions, we determine the initial and final *instantaneous* velocities.

Instantaneous Acceleration. If we choose Δt small enough

$$a = \frac{v_2 - v_1}{\Delta t} = \frac{\Delta v}{\Delta t} \tag{2.4}$$

will assume the meaning of **instantaneous acceleration,** since the small chosen time interval Δt can be made small enough to define an "instant," *i.e.*, a "point in time," as nearly as we desire.

The unit of acceleration is equal to the velocity unit over the unit of time, for instance,

$$\frac{\text{unit of velocity}}{\text{unit of time}} = \frac{\text{cm/sec}}{\text{sec}} = \frac{\text{ft/min}}{\text{sec}} = \frac{\text{mi/hr}}{\text{years}}, \text{ etc.}$$

A term such as $\frac{\text{cm/sec}}{\text{sec}}$ is frequently abbreviated for convenience by writing cm/sec^2.

Physics before Galileo

The quantitative concept of acceleration was developed and introduced into physics by Galileo in the course of his investigation of free fall. This work can be considered a turning point not merely in the history of physics but in the history of the natural sciences as well. Theretofore, people were attempting to philosophize laws of nature out of some vague, often meaningless "principles." They often drew perfectly logical conclusions from wrong premises and "derived" "natural laws" which were in contradiction to experience. Since no one made efforts to check these "laws" by experiment, they were accepted and dominated "scientific thought" for well over a millenium.

Aristotle's assertion that "bodies fall faster in proportion to their weight" provides a classic example of such "laws of nature." This statement is wrong, and Aristotle could have easily disproved it by trial. But the experimental method, which appears natural to us, evidently was not so obvious to him or to most of his contemporaries and followers. Aristotle's contributions to science and philosophy are monumental in fields involving pure reasoning and observation. They are for the most part worthless or lacking in fields such as physics where the acquisition of knowledge must depend on experimentation. In fact, his authority in physics dominated thought for many centuries and materially blocked progress in this field.

Galileo's Methodological Contribution

Galileo had the temerity to question Aristotle's "law of falling bodies." He argued: If a is a light stone and A a heavy one, then according to Aristotle A should fall faster than a. If I now tie them together to form a double stone, in falling, a should retard A, since it tends to fall more slowly, and the combination would fall faster than a but more slowly than A, *but* according to Aristotle, the double body $(A + a)$ is heavier than A and hence should fall faster![†] There was but one way out of this paradox, and this was to check experimentally Aristotle's questionable assertion that heavier bodies fall faster. This can be done simply by dropping a and A simultaneously and listening to observe whether they strike the floor at the same time. This experiment was well within the reach of Aristotle's laboratory facilities. Galileo performed experiments and discovered that all bodies, regardless of weight, would fall equally fast if not impeded by air resistance. In many cases air resistance is not so great as to prevent one from disproving Aristotle by free-fall experiments in air.

Having shown that Aristotle was wrong and that, except for air resistance, a feather and an ax, if dropped together, would fall side by side until they hit the ground, Galileo proceeded to ask, *how* does a body fall? When one drops a stone, how can one predict where it will be after one, two, three, or any number of seconds? And how does its velocity vary with time? His approach has become the classic pattern of physical research and will be discussed here in detail.

He started out with the common experience that a stone dropped from a height falls faster and faster: A stone dropped from a height of 50 ft hurts us more than one dropped from 5 ft, just as it is more painful to be struck by a fast-moving object than by a slowly moving one. *Evidently the falling motion is an accelerated one.* (1) Is the acceleration constant, does it (2) increase or (3) decrease as time goes on?

Galileo began with the first assumption of a constant acceleration as the simplest one, guided by his belief in simplicity of nature, and subjected it to experimental test.[†] (The practical virtue of adopting tentatively the simplest assumption to begin with consists in the fact that it is usually the simplest to prove or disprove. One should, however, be ready to drop it and to proceed with the investigation of less simple assumptions should the original one prove to contradict our experience.)

[†] Galileo's belief that Artistotle's statement was logically inconsistent was fallacious, since Aristotle made no statements about the behavior of parts of a body. This is one of the many instances in physics where an invalid argument proved fruitful by stimulating further investigation of a subject.

Since he had no direct means of measuring acceleration, he had to use an indirect method of approach, which is now commonly used in physics: You draw logical conclusions from your basic assumptions attempting to derive relationships between observable quantities. If the derived relationships are in contradiction to experiment, your basic assumptions have been proved wrong. In case of agreement, you are justified in accepting them tentatively as correct.

The observable quantities which Galileo could measure were time and distance. Thus, his preliminary question was: What relationship must exist between the distance traversed by a falling body and the time of fall if I assume the acceleration to remain constant? This question suggests the following line of action:

Derive the relation[‡] $s = f(t)$; then measure actual times and distances covered by falling bodies and see if they follow the theoretically derived law based on the assumption of constant acceleration.

Derivation of Galileo's Equations for Uniformly Accelerated Motion. Transform the definition equation for average acceleration,

$$\bar{a} = \frac{v_2 - v_1}{t} \tag{2.3}$$

by solving for v_2:

$$v_2 = v_1 + at$$

Since v_2 is the final velocity and v_1 the initial velocity, we relabel them for convenience: $v_2 = v_f$, and $v_1 = v_0$; hence,

Galileo's first "law":

$$\boxed{v_f = v_0 + at} \tag{2.5}$$

This equation enables us to find the final velocity v_f after any time t of falling (or moving) in **uniformly accelerated motion** (motion subject to a *constant* acceleration).

[†] He writes: "In the investigation of naturally accelerated motion we were led, by hand, as it were, in following the habit and custom of Nature herself, in all her various other processes, to employ only those means which are most common, simple and easy. . . . "

[‡] Read s is a function of t. This symbolism states that the distance traversed varies in a definite fashion with the time according to a certain formula which is not specified here. This symbolism is commonly used where the functional relationship is not known or is not stated for the sake of brevity.

But this is not the final equation we want. We are looking for $s = f(t)$. Hence, we have to introduce s from some relation between s, v, and t [$\bar{v} = s/t$, equation (2.1)]. The following consideration leads to the final formula: If a man's daily income is increasing uniformly throughout the year, it will be represented by a straight ascending line if we plot his daily income against time. From the data presented in Fig. 2.2, his daily average will be obtained as $(10 + 100)/2 = \$55$ per day.

Similarly, in the case of a uniform acceleration the growing instantaneous velocity will be represented by a graph resembling Fig. 2.2. If we replace dollars by centimeters per second, the average speed value will be given by $\bar{v} = (10 + 100)/2 = 55$ cm/sec, or in general terms

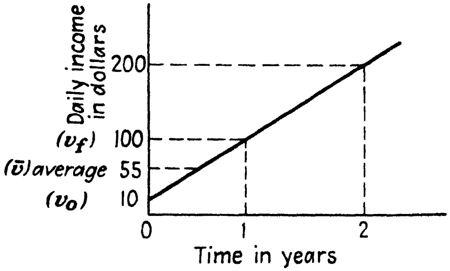


FIG. 2.2. Determination of average income in the case of a uniformly increasing income.

$$\bar{v} = \frac{v_0 + v_f}{2} = \frac{1}{2}(v_f + v_0) \tag{2.6}$$

since $v_0 = 10$ cm/sec and $v_f = 100$ cm/sec were the velocities at the beginning and at the end of the year. Substituting (2.5) in (2.6) we get

$$\underline{\bar{v}} = \tfrac{1}{2}[(v_0 + at) + v_0] = \tfrac{1}{2}(2v_0 + at) = \underline{v_0 + \tfrac{1}{2}at} \tag{2.7}$$

Substituting (2.1) in (2.7) we obtain

$$\frac{s}{t} = v_0 + \frac{1}{2}at$$

or Galileo's second "law":

$$\boxed{s = v_0t + \tfrac{1}{2}at^2} \tag{2.8}$$

This is the desired relation, $s = f(t)$.

If the body starts from rest, $v_0 = 0$ and the equation is simplified:

$$\boxed{s = \tfrac{1}{2}at^2} \tag{2.8a}$$

In order to see whether this relation between s and t actually is the law according to which bodies fall, Galileo could have proceeded according to the following plan: Drop the stone, and measure s after 1 sec; then $a = 2s$; thus, by measuring s, you get the value of a. Then measure t for various values of s, and see if the observed value of s for a given t always agrees with the value computed from formula (2.8a).

Galileo's Induction of the Laws of Free Fall. Galileo had no quick-acting stop watch to perform such tests on freely falling bodies. He approached this experimental problem indirectly. He realized that balls rolling or sliding down an inclined plane were performing a motion slower than but similar to free fall. In fact, he considered free fall as a special case of sliding down a vertical "inclined" plane. He confirmed the validity of equation (2.8) for balls rolling down a gentle slope and extrapolated its validity for steeper slopes than he could investigate and for the special case of free fall as well, *i.e.*, a slope of 90°. It should be emphasized that, although he never tested equation (2.8) for free fall directly, yet he affirmed the validity of his laws of uniformly accelerated motion for freely falling bodies and asserted that the acceleration of gravity is the same for all bodies on earth. (To be more precise, we should add: "when tested at the same locality.")

Equations (2.5) and (2.8) have been arrived at without resorting to experimentation. They are not laws of nature but rather definition equations of uniformly accelerated motion.

The assertion, however, that freely falling bodies move in accordance with equations (2.5) and (2.8) does constitute a law of nature and is equivalent to the statement that free fall is a uniformly accelerated motion.†

The special value of the acceleration due to gravity is designated by g. ($g = 980\ \text{cm/sec}^2 = 32\ \text{ft/sec}^2$, approximately.) g varies with geographic location and altitude.

In applying formulas (2.5) and (2.8) to free fall, replace a by g.

The Processes of Induction and Deduction

The process illustrated above of extrapolating beyond one's actual experience, of inferring a general law from selected observations of a few special cases, is called *induction*. It is by this process of reasoning that all general laws of nature are arrived at. We cannot possibly test all conceivable cases. The process of induction involves as a major premise the tacit assumption of the validity of the **postulate of the uniformity of nature**: *What happens to some members of a class of objects under certain*

---

† The same equations also hold for motion of bodies rolling or sliding down an inclined plane. In this case $a < g$.

conditions at certain times will always happen under similar conditions to all members of this class.

After establishing a general law by the *inductive process,* the scientist is enabled to deduce from the general law what would happen to a given object under certain specified circumstances. *Deduction* consists of an inference from the general law to the special case. For example, the computation of the position of a falling body at a given time from equation (2.8) is an example of a deduction. The solution of problems in physics will provide ample exercise in the practice of the deductive process.

Induction and deduction are analogous to the mathematical processes of extrapolation and interpolation.

Laws and Experience

All natural laws are derived from experience by induction. They enable us to make predictions of what would happen in some special cases. We cannot reason out laws a priori, *i.e.*, without previous experience.

Example

An antiaircraft projectile is fired vertically upward and reaches the height of $s = 40{,}000$ ft. (1) How long does it rise? ($t = ?$) (2) How long does it remain in the air? ($T = ?$) (3) What was its initial velocity? ($v_0 = ?$).

Solution. (1 and 3) We tabulate the given and the unknown data: $s = 40{,}000$ ft; $a = g = 32$ ft/sec²; $v_f = 0$. (The value of the final velocity, $v_f = 0$, at the highest point $s = 40{,}000$ ft above ground is given implicitly. We know that at this point the projectile stops and reverses its direction of motion.)

$$\boxed{t = ?}$$

To find t, we must obviously set up an equation which contains, on the one hand, the unknown, and, on the other hand, the given quantities s, g, and v_f. Such a relationship is provided by equation (2.8): $s = \uparrow v_0 t + \downarrow \frac{1}{2} a t^2$. Substituting the given data and remembering that the acceleration vector points downward whereas the v_0 vector points upward, we write (assigning a negative sign to the downward direction)

$$s = 40{,}000 = v_0 t - \tfrac{32}{2} t^2 \qquad (a)$$

This is an equation with two unknowns. We need a second equation containing one of the two unknowns or both of them to solve the problem. Such a second equation is provided by equation (2.5): $v_f = \uparrow v_0 + \downarrow at$. Substituting the given data into this equation we obtain

$$v_f = 0 = v_0 - 32t \qquad (b)$$

With the setting up of the two physical relationships, equations (*a*) and (*b*), the problem is solved as far as physical considerations are concerned. The rest is purely mathematical manipulations.

We solve equation (*b*) for t:

$$t = \frac{v_0}{32} \tag{b'}$$

Substituting equation (b') in equation (a) we get

$$40{,}000 = v_0\left(\frac{v_0}{32}\right) - \frac{32}{2}\left(\frac{v_0}{32}\right)^2 = \frac{v_0^2}{32} - \frac{1}{2}\frac{v_0^2}{32} = \frac{v_0^2}{64} \tag{c}$$

Hence, $v_0^2 = 64(40{,}000)$ or $\underline{v_0} = 8(200) = \underline{1600 \text{ ft/sec}}$

Substituting this value of v_0 in equation (b'), we get

$$\underline{t} = \frac{1{,}600}{32} = \underline{50 \text{ sec}}$$

2. The total time of flight T is equal to the time of rise t plus the time of fall t_F. To find the time of fall from the top height $s = 40{,}000$ ft, we can again use equation (2.8):

$$s = v_0 t_F + \tfrac{1}{2} g t_F^2$$

Since the initial velocity at the highest point is $v_0 = 0$, this equation is simplified as follows:

$$s = 40{,}000 = \tfrac{1}{2} \times 32 t_F^2$$

Hence, $$t_F^2 = \frac{40{,}000}{16} = 2{,}500$$

or $$\underline{t_F} = \underline{50 \text{ sec}}$$

We see that the time of fall is equal to the time of rise. The total time of flight is

$$\underline{T} = t + t_F = 50 + 50 = \underline{100 \text{ sec}}$$

QUESTIONS AND PROBLEMS

1. What experimental difficulties did Galileo encounter in his studies of free fall, and how did he overcome them?

2. How could you go about proving by simple means that all bodies fall alike?

3. Average velocity $= \dfrac{v_{\text{final}} + v_{\text{initial}}}{2}$. In which of the following cases does the above formula hold true?

a. The instantaneous acceleration $= \dfrac{\Delta v}{\Delta t}$.

b. The acceleration is constant.

c. The general case of varying velocity.

d. All cases by definition.

e. $a = \dfrac{s}{2} t^2$

4. In the following table, V represents the final velocity after t seconds.

t:	1	2	3	4	sec
V:	5	8	11	14	cm/sec

The motion in question could be (*a*) uniform; (*b*) uniformly accelerated, starting from rest; (*c*) uniformly accelerated with an initial velocity different from zero; (*d*) nonuniformly accelerated, the acceleration being proportional to time.

5. Can a body have a constant speed and yet have a varying velocity? Can a body have a constant velocity and yet have a varying speed?

6. (*a*) Is it possible for a car to be moving eastward while being subjected to a westward acceleration? (*b*) Can an object move westward with an acceleration which is neither eastward nor westward nor northward nor southward? Give some familiar examples.

7. A passenger traveling in a car with the speed of 10 m/sec encounters a train 200 m long moving in the opposite direction with a speed of 5 m/sec. How long will it take the passenger to pass the train?

8. Two bicyclists start moving toward each other with a uniform ground speed of 20 ft/sec along a straight path. Their initial distance apart is 800 ft. As soon as the first cyclist starts moving, a fly takes off from his forehead and flies at 100 ft/sec to the forehead of the second cyclist, lands there, and flies right back to the forehead of the first cyclist and so on, with constant speed until the two travelers collide and the fly is crushed between their foreheads. During this process the distance the fly has to traverse is getting shorter and shorter.

a. How long does the fly remain in the air? (Neglect the parking time on the foreheads.)

b. How great a distance does the fly traverse?

9. A professor of physics conducts experiments on free fall. He drops stones from the roof of his laboratory 64 ft above ground, starting his stop watch when the stone is released and stopping it when he hears the impact of the stone upon the ground. On one occasion, by mistake, he drops the stop watch while retaining the stone in his hand. Assuming that the stop watch was started at the moment of release and was merely stopped by the impact, what will it read?

10. A car traveling with the speed of 15 mi/hr is stopped in 5 sec. Express its average acceleration (*a*) in miles per hour per second; (*b*) in feet per second per second.

11. A car traveling with the speed of 80 ft/sec is being retarded at the rate of 20 ft/sec^2.

a. How fast will it move at the end of 2 sec?

b. How far will it have traveled during those 2 sec?

c. How far will the car travel before it stops?

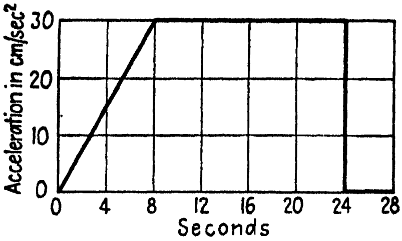
12. The following diagrams describe the motion of an automobile over a period of time: (*a*) acceleration, (*v*) velocity, and (*s*) distance from the starting point. Answer the following questions by analyzing these graphs:

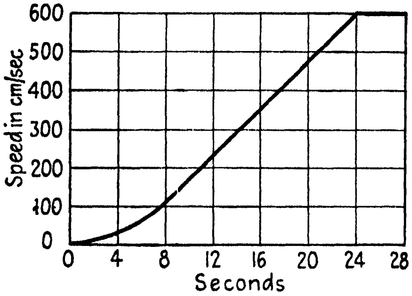
a. How would you describe the type of motion which the car performs during the first 8 sec?

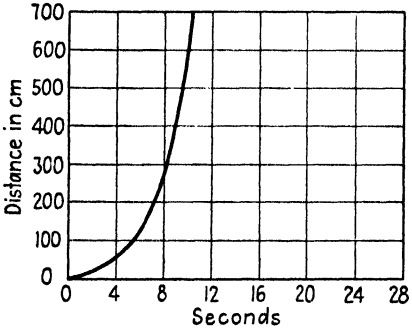
b. Then how would you designate the type of motion during the following 16 sec?

c. How does the car move after a total of 24 sec?

d. 8 sec after the start (1) how far has the car moved? (2) how fast is it moving?







13. You stand on a roof and throw a ball upward. You fail to catch it, and the ball falls to the pavement. As the ball passes the edge of the roof, its downward velocity is 30 ft/sec.

a. How fast will it move 2 sec after passing the edge of the roof?

b. If the ball strikes the ground at the end of this 2-sec interval, how tall is the building?

14. A baseball is knocked vertically upward to a height of 256 ft. Find (*a*) how long it moves upward; (*b*) total time the ball remains in the air; (*c*) the ratio between the time of upward and that of downward motion; (*d*) the initial speed of the ball; (*e*) the speed of the ball at the highest point.

15. How high do you have to throw a ball in order to catch it 2 sec later?

16. If you drop a body from a high cliff ($v_0 = 0$) and 2 sec later release another body, (*a*) how will the relative distance between the two bodies change as they fall? (*b*) how will the relative velocity change?

17. In broad jumping, does it matter how high you jump? What factors determine the span of the jump?

CHAPTER 3

KINEMATICS: 2. ANGULAR MOTION AND CURVILINEAR TRAJECTORIES

a. In what units would you measure the rate of escape of gas from a tank?

b. Is the escape of gas from a tank an accelerated, retarded, or uniform process?

c. If a wheel makes 3 revolutions per second, how large an angle (1) in degrees, (2) in angular measure, does a spoke sweep in *(a)* $\frac{1}{8}$ sec, *(b)* 2 sec?

d. How could you experimentally produce uniformly accelerated rotation?

e. Can the same formula (identical in form and letters) be used to describe two different physical processes?

f. Does a ball you drop in (1) a speeding train, (2) a descending elevator fall as fast as one released from a tower?

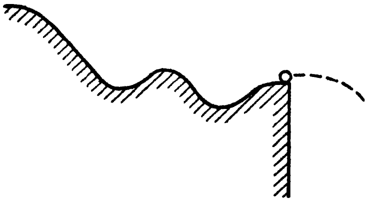
g. What kind of curves (approximately) do you see when you watch (1) bursting fireworks? (2) water escaping from a fire hose?

h. Can a body move in two different directions at the same time?

i. Does it make sense to ascribe to a body a uniform and an accelerated motion at the same time?

j. What is the relation or distinction between a physical law and a physical formula?

k. A ball slides down a smooth curved groove and continues to move freely after it leaves the groove. How does the shape of the groove influence the shape of the trajectory?



l. A wheel rolls, without slipping, along the floor at 10 cm/sec. The radius of the wheel is 5 cm. What is the angular velocity of the wheel? What is the instantaneous velocity of the top of the wheel relative to the floor? Of the bottom of the wheel relative to the floor?

Generalization of the Velocity Concept

The concept of velocity can be generalized. For instance, instead of determining how many centimeters a body would move each second, we can determine how many gallons of water escape each minute through a faucet and speak of a velocity of efflux (outflow), or we can note the angle through which a wheel would rotate in a second and designate this quantity as the angular velocity of rotation, etc.

Rotational Motion. We shall analyze as an example the case of rotation of a wheel about a fixed axis. Just as in the case of linear motion, rotation may be uniform or accelerated. The rotational displacement is measured against a fixed external mark *m* by the angle θ (Fig. 3.1).

The angular velocity ω is defined as change of angular position (change of orientation) per unit time:

$$\bar{\omega} = \frac{\theta_2 - \theta_1}{t} = \frac{\theta_f - \theta_0}{t_2 - t_1} = \frac{\Delta\theta}{\Delta t} \tag{3.1}$$

the angular acceleration α as change of angular velocity per unit time:

$$\bar{\alpha} = \frac{\omega_2 - \omega_1}{t} = \frac{\omega_f - \omega_0}{t_2 - t_1} = \frac{\Delta\omega}{\Delta t} \tag{3.2}$$

The equations of uniformly accelerated rotational motion can be derived from those of uniformly accelerated linear motion by the following simple consideration:

If we attach a weight W to a string wound about the periphery of a wheel as shown in Fig. 3.1, the weight will descend with uniform acceleration, forcing the wheel to rotate. We shall show now that the resulting rotation will be a uniformly accelerated one (*i.e.*, will have a constant angular acceleration α) and will obey formulas identical in form to those for uniformly accelerated linear motion [(2.5) and (2.8)].

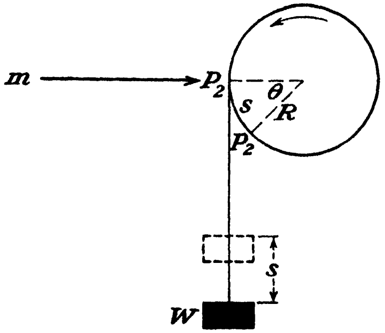


Fig. 3.1. Relation between linear and angular displacement.

Preliminary Consideration. Angles can be measured in **degrees** (the periphery of a circle being divided into 360 equal parts) or in **angular measure (radians)** by the ratio of the arc cut off on the periphery of a circle to its radius:

$$\theta = \frac{s}{R} \text{ radians} \qquad \text{(Fig. 3.1)}$$

or

$$\boxed{s = R\theta} \tag{3.3}$$

From this and the equations (3.1) and (3.2) follow the relations

$$\underline{v} = \frac{s}{t} = R\frac{\theta}{t} = \underline{R\omega} \qquad \left(\text{where } \frac{\theta}{t} = \omega\right) \tag{3.4}$$

and

$$\underline{a} = \frac{v_2 - v_1}{t} = \frac{R\omega_2 - R\omega_1}{t} = R\frac{\omega_2 - \omega_1}{t} = \underline{R\alpha} \tag{3.5}$$

The descent of the weight W is given by the amount of string unwound when the wheel rotates through the angle θ. This, in turn, is equal to the length of the arc s.

Thus, we can apply the formulas for linear accelerated motion [(2.5) and (2.8)] to the descent of W and express s, v, and a in those equations in terms of θ, ω, and α.

$$v_f = v_0 + at \tag{2.5}$$

(3.4) and (3.5) in (2.5):
$$R\omega_f = R\omega_0 + R\alpha t \tag{3.6}$$

dividing by R,
$$\boxed{\omega_f = \omega_0 + \alpha t} \tag{3.6a}$$

similarly,
$$s = v_0t + \tfrac{1}{2}at^2 \tag{2.8}$$

(3.3) to (3.5) in (2.8):
$$R\theta = R\omega_0t + \tfrac{1}{2}R\alpha t^2 \tag{3.7}$$

dividing by R,
$$\boxed{\theta = \omega_0t + \tfrac{1}{2}\alpha t^2} \tag{3.7a}$$

The formal analogy between the equations for linear and angular motion is obvious. We merely have to make the substitutions listed in the following column to transform a linear equation into one applicable to rotational motion:

Linear quantities	*Angular quantities*
s	θ
v	ω
a	α

Such formal analogies between equations are frequently encountered in different fields of physics. The analogy usually also extends to the methods of solution of problems involving the analogous laws.

Superposition of Motions. A bullet or a ball may be projected straight (1) vertically upward, (2) vertically downward, (3) horizontally, or (4) obliquely. What will its motion be like?

Cases 1 *and* 2. The ball will move along the vertical, and its position and velocity will be given at any instant by equations (2.8) and (2.5). When the ball is thrown downward, the initial velocity has the same direction as the gravitational acceleration which we indicate by choosing a positive sign for both:† $v_f = \downarrow v_0 + \downarrow gt$. When the ball is projected upward, its initial velocity has a direction opposite to the acceleration of gravity, which we indicate by choosing a negative sign for the former. Thus, a negative sign designates, according to our convention, upward motion, and a positive sign downward motion: $v_f = -\uparrow v_0 + \downarrow gt$. From this equation we can trace the behavior of the ball as time elapses: Initially ($t = 0$) the positive term is equal to zero, and $\underline{v = -v_0}$; our ball moves upward. As time goes on, the second term grows so that the sum on the right side approaches zero; *i.e.*, the upward motion slows down. At a certain moment when $gt = v_0$, the ball stands still and a moment later reverses its course and continues moving downward with ever-increasing speed, since the second term gt grows in proportion to the time and becomes finally predominant.

---

† We could have chosen the negative sign just as well.

We can describe the above process by saying that the ball actually performs two independent motions at the same time: (1) It moves upward with the uniform velocity v_0, and (2) it continues to fall at the same time with uniform acceleration. The resultant motion is a superposition of the two independent motions. This state of affairs may be clarified by the picture of a train speeding at constant velocity eastward while a passenger is running with uniform acceleration westward along the train. The passenger's accelerated motion is superposed upon the uniform motion of the train so that his motion relative to the ground is similar to the motion of the ball in the above example.

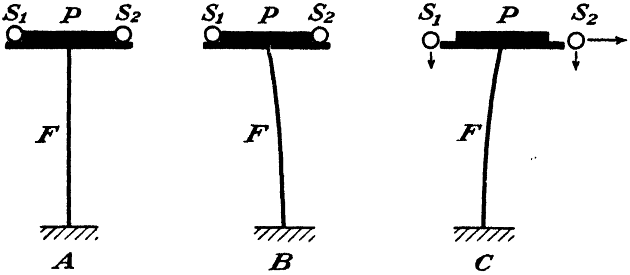


FIG. 3.2. Demonstration of "independence of motion."

Galileo found that the vertical accelerated motion of a falling body is not influenced by any other motion which it may perform at the same time. For example, a stone thrown upward in a uniformly moving train or in a uniformly ascending or descending elevator moves relative to the train or elevator exactly as it would move relative to firm ground if thrown similarly by an observer at rest with respect to the ground.

This independence of motion was demonstrated by Galileo with the aid of an apparatus similar to the one shown in Fig. 3.2. Two identical spheres S_1 and S_2 are resting at the ends of a platform P which is supported by an elastic spring F (Fig. 3.2A). In B the platform is shown pulled slightly to the left. When it is released, it snaps to the right due to the action of the spring F. As a result of this impulse, the sphere S_2 is projected horizontally to the right (Fig. 3.2C), while the support is withdrawn from under S_1, which begins to fall freely. For practical purposes one can consider the release of the spheres S_1 and S_2 to be simultaneous. If the falling motion of S_2 is not affected by its horizontal motion, one would expect the spheres S_1 and S_2 to be at the same height at any moment. The correctness of this assumption is demonstrated by the fact that the two spheres strike the table simultaneously.

Projectiles. *Case* 3. Let us now examine the shape of the trajectory described by the sphere S_2 which moves uniformly to the right as it falls freely (Fig. 3.2). To find the position of S_2 at any time, we have to ask two questions: How far did the uniform motion carry our body along the horizontal? How far did it fall during this time? This will give us the instantaneous position of the projectile. For convenience we imagine these two displacements, which actually occur simultaneously, to be taking place in two successive steps. And by choosing different values of t, we can construct the curved path of a horizontally projected body point by point. It turns out to be a parabola, a curve whose points are determined by the following relationship: The vertical distance y of a point below the point of origin is proportional to the square of its horizontal distance x away from the origin. This relation can be easily seen from Fig. 3.3. The horizontal displacement x is determined by the uniform velocity v_x of horizontal motion:

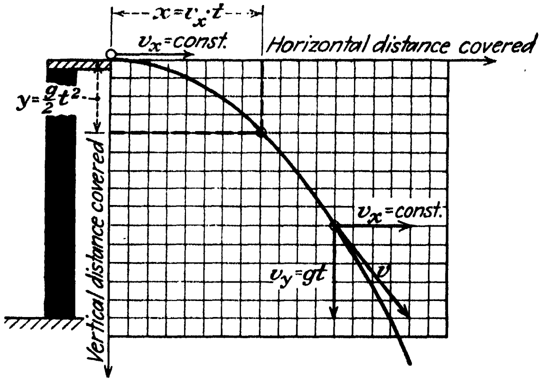


FIG. 3.3. Construction of the path of a projectile.

$$x = v_x t \qquad (a)$$

the vertical displacement due to free fall is given by

$$y = \frac{g}{2} t^2 \qquad (b)$$

Solving (*a*) for t, we get $t = x/v_x$. Substituting this value for t in equation (*b*), we obtain $y = [(g/2)\,(1/v_x^2)]x^2$. Since the terms in the brackets $g/2v_x^2 = c$ are constant, the equation $y = cx^2$ determines a parabola.

We encounter such parabolic trajectories not only in mechanics in connection with artillery, hurling of stones, liquid jets escaping from a hose, and jumping animals but also, as we shall see later, in connection with moving electrons in oscilloscopes and isotopic atoms in mass spectrographs and isotope separators.

Case 4. The trajectory illustrated in Fig. 3.3 is the special case of a parabola obtained with a horizontal initial velocity. If the initial uniform velocity v_x were directed obliquely upward, we should still get parabolas but their shape would be as shown in Fig. 3.4. These curves can be obtained quite simply by projecting an ink-coated sphere S by means of a spring cannon C up a gently inclined drawing board. As a result of the uniform speed v_x acquired from the cannon and the uniform acceleration which carries the sphere down the gentle slope of the drawing board, the sphere traces a parabola on the paper which is indicated in Fig. 3.4 for two cases of different directions of v_x.

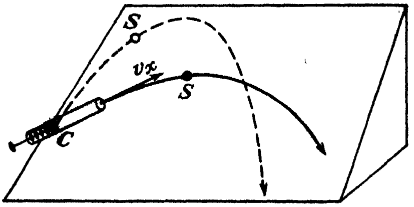


FIG. 3.4. Demonstration of the parabolic trajectory in oblique projection.

On the Use of Formulas

Physical formulas such as (2.5), (2.8), etc., are a shorthand notation of physical laws. The symbolic representation of physical quantities and the mathematical treatment of physical problems are, contrary to some laymen's belief, not complicating features which make physics such a difficult subject, but rather time- and thought-saving devices which ease the mental task of the physicist, just as a tool or a machine eases the physical task of a laborer. Physical formulas are a clear testimony to the cumulative nature of science. By using them we make the brains of Newton, Leibnitz, Maxwell, etc., collaborate in the solution of our problems.

It is not at all a "stupid mechanical process" to solve problems by appropriate manipulation of such formulas as long as we recall the origin and the physical meaning of the laws they symbolize. It is largely due to the use of this time- and effort-saving mental device that the exact natural sciences have advanced as far as they have.

Example 1

A hoop of radius $R = 50$ cm rolls downhill with the constant acceleration of $a = 25$ cm/sec². How many revolutions does the hoop make in $t = 10$ sec starting from rest?

Solution. We treat this as a problem in angular motion. The angular acceleration is given by equation (3.5), $\alpha = a/R = \frac{25}{50} = \frac{1}{2}$ radian/sec². The angle covered in t seconds is given by equation (3.7a): $\theta = \omega_0 t + \frac{1}{2}\alpha t^2$. Since the hoop starts from rest we have $\omega_0 = 0$ and $\theta = \frac{1}{2}\alpha t^2 = \frac{1}{2} \times \frac{1}{2} \times (10)^2 = \frac{100}{4} = 25$ radians. Now, 1 revolution corresponds to 2π radians; hence, the wheel makes

$$\underline{N = \frac{25}{2\pi} \simeq 4 \text{ revolutions}}$$

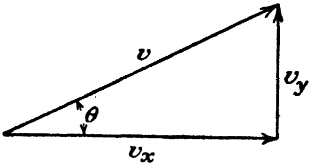
Example 2

A heavy stone is projected from a small slingshot with a speed of $v = 12$ ft/sec aiming at an angle of $\theta = 30°$ against the horizontal. The stone hits a target which is as high above the ground as the slingshot. How far is the target away from the slingshot? (Neglect air friction.)

Solution. If we knew the time T during which the stone remains in the air and the horizontal component v_x of its motion, we should know how far from the slingshot the stone will be after T seconds. This distance is given by $x = v_xT$.

The time T is equal to the time of rise plus the time of fall, or since both are equal (negelecting air friction), $T = 2t$ where t is the time of rise.

We proceed therefore as follows: (1) Find v_x, (2) find $T = 2t$, and finally (3) find $x = v_xT$.



1. We resolve the initial velocity into a vertical upward component v_y and a horizontal component v_x. From the diagram,

$$\sin\theta = \frac{v_y}{v}; \qquad v_y = v \sin 30 = 12 \times \frac{1}{2} = 6 \text{ ft/sec}$$

and

$$\cos\theta = \frac{v_x}{v}; \qquad \underline{v_x} = v \cos 30 = 12 \times \frac{1}{2}\sqrt{3} = \underline{10.4 \text{ ft/sec}}$$

2. We now disregard the horizontal motion and consider the vertical upward motion of the stone. The initial upward velocity is $v_0 = v_y = 6$ ft/sec; hence, using equation (2.5) and remembering that $v_f = 0$ at the highest point, we get

$$v_f = 0 = \uparrow v_0 + \downarrow at = 6 - 32t$$

or

$$t = \tfrac{6}{32} = 0.187 \text{ sec}$$

Hence, the time of flight is $\underline{T = 2t = 0.374 \text{ sec}}$

3. From 1 and 2 follows the "range" $\underline{x} = v_xT = 10.4(0.374) = \underline{3.90 \text{ ft.}}$

QUESTIONS AND PROBLEMS

1. What is the ratio of the angular velocity of the hour hand to that of the minute hand of a clock?

2. Compute the angular velocity of the rotation of the earth about its axis (*a*) in degrees per second, (*b*) in radians per second.

3. Compute the relative linear velocity between a phonograph needle and a recording disk spinning at the rate of 78 rpm at a point 3 in. from the axis of rotation.

4. A bicycle moving with the speed of 1 m/sec stops within 5 sec. Assuming constant acceleration and a wheel radius of $R = 40$ cm, compute (*a*) the angular acceleration of the wheel, (*b*) how many revolutions the wheel makes before it stops.

5. A weight is attached by means of a string to a wheel as illustrated in Fig. 3.1. The wheel is rotating with an angular acceleration of $\frac{1}{10}$ radian/sec^2. How far does the weight descend in 4 sec starting from rest if the diameter of the wheel is 10 cm? How fast does it move 4 sec after starting?

6. If you want to shoot as high as possible, in what direction should you aim the rifle?

7. What is the shape of the path of a jumping kangaroo?

8. In sharpshooting, how should you aim the rifle in relation to the target?

9. How would the procedure in Prob. 8 be modified if the target were released to fall freely at the moment the bullet leaves the muzzle?

10. A hawk flying uniformly at 60 mi/hr in a straight line at a height of 800 ft drops its prey. How far will the bird be ahead of the prey when the latter strikes the ground? (Neglect air resistance.)

11. An airplane traveling horizontally at 300 mi/hr is shot down over a certain village. The airplane is falling for 16 sec.

a. How far from the village does it fall?

b. How high was it above the ground when hit? (Neglect air friction.)

12. A projectile is fired at an angle of 30° with respect to the horizontal level with an initial speed of 1,200 m/sec.

a. What curve does the projectile describe?

b. How high does it rise?

c. How long does it remain in the air?

d. How far from the cannon does it hit the ground? (Assume the cannon to be on ground level.)

CHAPTER 4

DYNAMICS: 1. NEWTON'S LAWS (FIRST AND SECOND)

a. How is the field of mechanics subdivided?

b. Can you think of a simple way of measuring forces?

c. What is the relation between the concepts of weight and of force?

d. Is a force necessary to keep a body moving?

e. Which of Newton's laws defines the concept of a force qualitatively?

f. Which of Newton's laws defines the concept of a force quantitatively?

g. What is an "unbalanced force"?

h. To what verbal statements are the following shorthand notations equivalent?

$$F = ma; \qquad m = \frac{F}{a}; \qquad \frac{F}{ma} = 1$$

i. How could you measure forces with a yardstick and stop watch?

j. What is mass? How can you measure it?

k. Which two properties of a body are measured by its mass?

l. What is density? In what units do you measure it?

m. Can a body have mass and yet no weight?

n. Can a body have weight and yet no mass?

o. Does a spring scale measure weight or mass?

p. Does an equal-arm balance measure weight or mass?

q. If you are given the mass of a body, by what equation could you determine its weight?

r. Are Newton's first two laws laws of nature or definitions?

s. Would our physical knowledge suffer if we were to ignore Newton's first law?

t. Is the value of g the same everywhere in the universe?

u. Why is g independent of weight?

v. What is an axiom? What is a postulate? Can one axiom be derived from another?

Division of Mechanics into Kinematics and Dynamics

In the previous sections we discussed **kinematics,** which is limited to the description of the motion of bodies without raising the question of why they move as they do. It is the object of **dynamics** to account for the peculiarities of observed motions by ascribing them to the action of forces.

Galileo's Law of Inertia. The intuitive concept of a force must have originated very early in human history. We usually say that we have to exert a force in order to lift a weight or to hold it in suspension or to stretch a spring. Is there any connection between this agent (force), which is capable of neutralizing gravity, and the motion of bodies?

Is a force perhaps required to keep a body in motion? Aristotle believed so, and so did his followers until Galileo's time. Galileo concluded from his experiments in which he allowed spheres to roll down an inclined plane and to continue rolling along a horizontal plane that the only reason why the spheres stop eventually is a frictional counterforce. He based his conclusion on the following observation: When a smooth sphere is released from point 1 of the inclined plane A (Fig. 4.1), it will roll up the inclined plane B almost to the same height H from which it was released (point 2). If the sphere and the surface of the inclined plane are polished more carefully, the sphere rolls up more nearly to the original height. He concluded that, in the limiting case of ideally smooth surfaces, the sphere would roll up precisely to the original height of release.

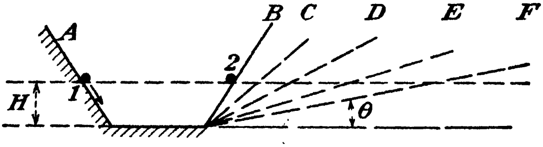


FIG. 4.1. Galileo's experiment with inclined planes.

He now proceeded to incline plane B more and more, making it assume successively positions C, D, E, F, etc. In all these cases his previous conclusion should apply; namely, the sphere would roll until it reached the height of release. But for a greater inclination of the plane B the point 2 will be located farther away from point 1. He concluded that for an angle of inclination of $\theta = 0°$ the sphere would roll to infinity, since it will not reach its original height at any finite distance from 1.

Furthermore, since deceleration of the sphere rolling up the plane B approaches zero as the plane approaches the horizontal position, the smooth sphere should proceed to infinity with a uniform speed in a straight line along a smooth horizontal plane. Ordinarily, this tendency of a body to move uniformly forever is masked, since the body usually is supported or otherwise in contact with another body, which exerts a retarding force of friction opposite to the direction of motion. We know from experience that to maintain the body in uniform motion we must pull or push it in the direction of motion in order to overcome friction.

Newton's First Law of Motion. Newton went further than Galileo in generalizing this induction. He adopted the equivalent of the following statement as a general law of motion (**Newton's first law of motion**):

"**A body in motion continues moving forever in a straight line with a constant velocity,** *unless acted upon by an external unbalanced force.*"

This statement goes beyond Galileo, since it is not restricted to rolling on a horizontal plane. It is meant to be quite generally applicable to any body upon which either no forces are acting or the acting forces cancel each other. It also applies to bodies at rest, since the state of rest is a special case of a uniform linear motion for $v = 0$:

"A body remains at rest, unless an external unbalanced force acts upon it."

This reluctance of material bodies to change spontaneously their original state of uniform linear motion is called **inertia** (meaning laziness).

If we define a force through its ability to lift a weight or to stretch a spring, this law appears as a daring induction, since no one has ever tested or ever will be able to test a moving body under such ideal conditions. Even far out in the universe a body will be attracted by the distant stars and, hence, could not be expected to move strictly uniformly; besides, the observer himself, who should be close enough to see the body, would attract it. However, we know that the more closely we approach the condition printed in italics in the preceding statement of the law, namely, the absence of an external unbalanced force, the more nearly does a moving body fulfill the first part of that statement.

A definition of the concept of force on the basis of gravity or distension of elastic bodies is, however, not satisfactory primarily because of our inability to preserve a permanent standard of a force so defined. The weight of a body at a specified locality may vary in the course of time owing to geologic changes, such as "uplift," and the force required to produce a specified elongation of a spring varies as the spring ages and is stretched repeatedly.

What Is a Force? We can avoid these difficulties and the difficulty of having to "prove" Newton's first law by making this law the basis of our definition of the concept of force. This approach to Newton's first law is suggested by an analysis of its statement.

Whereas the first part of Newton's first law expresses a natural tendency of isolated moving bodies, *the part printed in italics is of particular significance to us, since it can be interpreted as defining kinematically what we mean by a "force":* The force is that which is capable of preventing a body from moving uniformly in a straight line; *i.e.*, **a force is an agent capable of accelerating a body.**

(It required Newton's deep insight to realize that the same agent which can stretch a spring is also capable of accelerating a body.)

The acceleration is to be taken in its usual general sense: a change in magnitude or direction of the velocity or both. Even a mere deviation from the original course without a change in the magnitude of the velocity does imply acceleration!

This law is a purely *qualitative* definition of the term force, defining force as the cause of acceleration. It implies the possibility of measuring this unobservable something (force) by the observable acceleration of bodies which it produces. The interpretation of this law as a definition of the concept of force is the deeper reason why this law cannot be proved experimentally. One can neither prove nor disprove a definition.

Newton's Second Law of Motion. Newton has defined force as a quantity proportional to the acceleration which it produces, having the same direction as the acceleration.

Newton's second law of motion: $\boxed{F = ma}$ (4.1)

This second law of Newton constitutes a *quantitative* **operational definition of the concept of force.** (m stands for a proportionality constant which depends on the body used.)

Operational Definition of Force. Equation (4.1) indicates how, by measuring time and distance alone, one can measure forces: A body is selected as a standard to which we may arbitrarily assign the value of $m = 1$. The force then becomes numerically equal to the acceleration which it imparts to this standard body. Thus, *by measuring a (with yardstick and stop watch) one can measure forces.*

Inertial Mass. The proportionality constant† m of Newton's second law is called **mass.** Its physical meaning can be seen by rewriting equation (4.1):

$$a = \frac{F}{m} \tag{4.1a}$$

We see that an increase in mass would diminish the acceleration caused by a given force. In other words, an increase in mass renders the body less accelerable. Thus, mass appears to act as a "resistance to acceleration," or as is easy to recognize, it is associated with the property of matter which we call inertia.

But how can we increase the mass of a body? Suppose we are accelerating by a constant force a pail of water (Fig. 4.2). How can we render it less accelerable? By pouring in more water. In other words, to increase the mass we increase the quantity of matter in the pail. Newton designated *mass* as the *quantity of matter* and defined it as follows:

---

† We shall see in Chap. 40 that this original definition of m as a proportionality *constant* cannot be upheld in relativistic mechanics, where we deal with bodies moving at speeds close to the velocity of light.

$$\boxed{m = Vd} \tag{4.2}$$

where V is the volume of the body and d is its "density." The density is commonly defined nowadays as mass per unit volume. Some authors have accused Newton of circular reasoning, since he defined mass in terms of density, which we in turn define in terms of mass.†

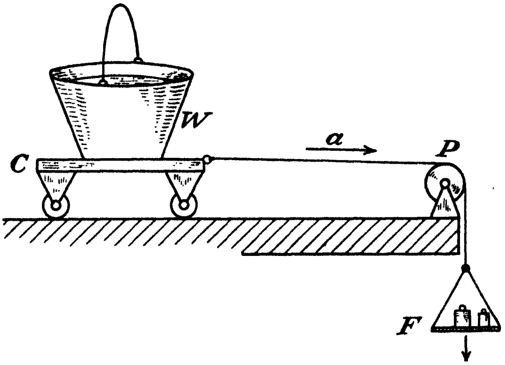


FIG. 4.2. Demonstration of acceleration of a variable mass by a constant force. Weights placed on the pan F provide the constant force which accelerates the car C on which is placed the pail of water W. The accelerated mass is varied by varying the amount of water in the pail.

Operational Definition of Mass. We see that mass, like time and distance, has remained an undefined concept. But just as these last two concepts lend themselves to an operational definition, mass can similarly be defined operationally; *i.e.*, we can state how masses could be measured by comparison with a standard mass, *e.g.*, standard kilogram preserved in Paris. We can compare masses by applying a constant force first to the standard and then to the unknown mass to be determined while measuring the resulting accelerations a and A in both cases. (Or we can apply the same force simultaneously to both masses by means of a compressed spring placed between them.) If the acceleration A imparted to the unknown mass is one-half as large as that of the standard mass, we know that the inertia of the unknown mass is twice as great as that of the standard. This follows from Newton's second law:

---

† It is likely that Newton did not define density as we do today but that he considered it as a primary intuitively evident concept not definable in terms of simpler ones.

For unknown mass: $F = MA$
For standard mass: $F = ma$
Ratio: $\frac{F}{F} = 1 = \frac{M}{m}\frac{A}{a}$ or $\frac{M}{m} = \frac{a}{A}$

from which the unknown mass can be expressed in terms of the standard: $M = (a/A)m$. This is obviously a cumbersome method of comparing masses. Is there a simpler way?

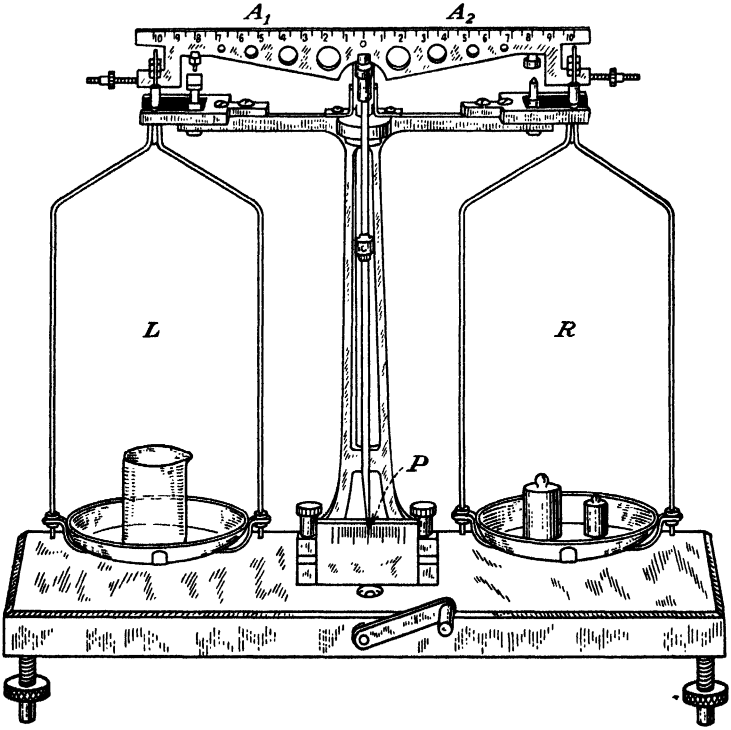


FIG. 4.3. A chemical balance. (*Courtesy of Central Scientific Company.*)

Gravitational Mass. In addition to inertia, matter possesses another property: **gravity**, or heaviness. All material bodies are observed to be attracted toward the center of the earth. Experiments show that gravity is proportional to inertia so that, for instance, a mass which is twice as inert as the standard is also twice as heavy. Thus, instead of comparing the inertia of masses we can compare their gravity, or **weight**, in the process of measuring mass.

This can be done conveniently by means of a balance (Fig. 4.3). The arms A_1 and A_2 of the balance are equal. The weights and hence masses on the two pans L and R of the scale are equal to each other when the horizontal balance arm remains in equilibrium, which is indicated by the central position of pointer P.

Relation between Weight and Inertial Mass. The fact that weight is proportional to inertia (mass) is expressed by the following equation, where g is a proportionality constant:

$$\underline{W = gm} \qquad (a)$$

On the other hand W is a force (the force of gravity, $W = F$) exerted upon the mass m, and Newton's second law should apply to it:

$$F = ma = \underline{am = W} \qquad (b)$$

Comparing (a) and (b), we see that $g = a$, that is, g has the meaning of an acceleration, namely, the acceleration which the grativational force of the earth (**weight**) imparts to the mass m.

Hence, the weight of any body can be obtained by multiplying its mass by the acceleration due to gravity:

$$\boxed{W = mg} \qquad (4.3)$$

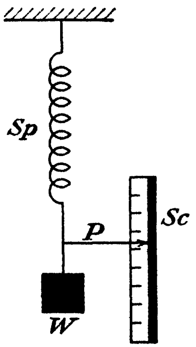


FIG. 4.4. A spring scale. Sp, spring; Sc, scale; P, pointer; W, weight.

In practice we can set $g = 32\ \text{ft/sec}^2 = 980\ \text{cm/sec}^2$, since, on the surface of the earth, all bodies fall with (roughly) the same acceleration.

We can see now that the balance of Fig. 4.3 compares weights as well as masses (as long as $g \neq 0$); its reading is independent of g, since a change in g affects the content of both pans equally. The reading of a spring scale (Fig. 4.4), however, depends on g as well as on m ($F = W = mg$).

If a balance shows that a body weighs 2 kg, it will give the same reading at the North Pole or on the moon. A spring scale, however, measures weight proper of a single mass and will give a varying reading according to geographic location and altitude above sea level.

Why Is g Independent of W? But if greater masses are acted upon (and hence accelerated) by greater gravitational forces, why do heavy and light bodies fall with the same acceleration? This is a simple consequence of Newton's second law: Since W is proportional to m [equation (4.3)], whenever you double W (and, hence, the accelerating force), you also double m, the "resistance to acceleration," so that the ratio $W/m = g$ remains constant and, hence, the acceleration does not change as W is altered.

Newton's Axiomatic System of Mechanics

In organizing his system of mechanics Newton appears to have adopted the classic pattern of the geometry of Euclid. Euclid begins with a set of definitions (some of which are only apparent definitions, actually leaving undefined the concept to be defined) and a number of mutually independent axioms and postulates expressed in terms of those concepts. These axioms and postulates are supposedly intuitively evident statements which cannot be proved but which nobody doubts. From his postulates he then proceeds to deduce logically all the theorems of geometry.

This orderly system appears worthy of emulation. If we follow Euclid's example, our ultimate aim will be to set up as small a number as possible of **axioms of physics** (we shall call them **fundamental laws**) and subsequently to try to deduce other physical laws from them.

Newton followed this pattern rather closely. His undefined concepts were distance, time, and mass, and from these he derived other concepts such as velocity, acceleration, force, etc. Finally, he set forth his **three fundamental laws of motion,** which are to mechanics what Euclid's axioms are to geometry:

1. **A body at rest will remain at rest and a body in motion will continue in motion with constant speed in a straight line as long as no unbalanced force acts on it.**
2. **Force equals mass multiplied by acceleration; its direction coincides with the direction of the acceleration.**
3. **To every action there is an equal and opposite reaction** (see Chap. 5).

Relation between Newton's First and Second Laws. If, in the equation $F = ma$ (Newton's second law), we set $F = 0$ (no unbalanced force acting), then $ma = 0$, and since m is not zero, $a = 0$, *i.e.*, there is no acceleration, which means the body moves straight ahead at a constant rate or remains at rest. (Compare with statements of Newton's first law.)

The second law can be considered as a quantitative operational definition of force, whereas the first law can be considered as a qualitative definition of force. It implies the assertion that only an unbalanced force (*and no other agent*) can be the cause of acceleration of a body. This assertion is not implied by the second law. Hence the first law is not merely a superfluous statement of a special case which is covered by the second law (namely, the case of $F = 0$).

APPENDIX I

Hooke's Law

There are different means of exerting a force upon a given body. For instance, we can subject it to the gravitational attraction of another body, such as the earth or the moon; we can expose it to the attraction of an electrified body; we can pull it by means of a rope or a wire; or we can push it by means of a rod. In all these cases the magnitude of the force depends on certain factors controlling the interaction between the bodies. In the case of gravitational interaction there is a law (Newton's law of gravitation, Chap. 13) which expresses the force of interaction in terms of the distance between the two bodies and of their masses. In the case of electrical interaction Coulomb's law (Chap. 14) determines the force exerted by one electrically charged body upon another electrified body in terms of their distance, electrical charge, and the so-called *dielectric constant* of the fluid in which the two bodies are submerged. Similarly, in the transmission of a force through a pulling wire or a pushing rod the force exerted upon the body can be expressed in terms of the deformation of the wire or the rod which transmits the force.

Whenever you accelerate a mass by pulling a wire to which it is attached, the wire is slightly elongated. Whenever you push a body by means of a rod, it can be shown by measurement that the rod is slightly contracted. If the contraction of the rod or the elongation of the wire is zero, the transmitted force is zero. Hooke found that for small elongations and contractions the transmitted force F is proportional to the change in length Δx:

$$\boxed{F = k\,\Delta x} \qquad \textbf{(Hooke's law)} \qquad (4.4)$$

The elastic constant $k = F/\Delta x$ has the meaning of the force which must be transmitted or applied to the rod or wire to produce unit deformation. In other words, $k = F$, when we set Δx equal to the unit of length.

This law applies practically to any rod or wire, regardless of the shape of the cross section, as long as the relative deformation $\Delta x/x$ is kept below a critical value (the elastic limit), where x is the original length of the undeformed wire (or rod) and Δx the change in length. When the force is removed, the wire or rod resumes the original length. Bodies which assume their original size and shape when the deforming force is discontinued are referred to as being **elastic.** A spring is a familiar device which obeys Hooke's law for large elongations or compressions.

Relationships similar to Hooke's law have been established for compression of solids, liquids, and gases. We shall consider them in Chap. 26 in connection with sound.

APPENDIX II

Units of Mass and Force

A. **The CGS System** (centimeter, gram, second)

Primary Units

Centimeter: Defined as $\frac{1}{100}$ of the international meter preserved in Paris.

Second: Defined as mean solar second or 1/86,400 of a mean solar day.

Gram: Unit of mass. 1 gm is 1/1,000 of a kilogram, which is defined as the mass of a standard platinum body preserved at the International Bureau of Weights and Measures. 1 gm is very nearly equal to the mass of 1 cm^3 of water at 4°C.

Derived Unit

Dyne: Unit of force. The definition of the unit of force is based on Newton's second law: $F = ma$. We specify: **We shall define that force as our unit (1 dyne) which will impart to a unit mass (1 gm) a unit acceleration (1 cm/sec²).** In other words, when we choose $m = 1$ gm and $a = 1\ \text{cm/sec}^2$, our equation will read by definition

$$F = (1\ \text{gm})(1\ \text{cm/sec}^2) = 1\ \text{dyne}$$

B. **The British Absolute System**

This is analogous to the cgs system.

Primary Units

Foot, second, pound: Unit of *mass*, defined as the mass of a preserved standard body.

Derived Unit

Poundal: Unit of force. 1 poundal is the force which will impart to a unit mass (1 lb) the unit acceleration of 1 ft/sec²:

$$F = (1\ \text{lb})(1\ \text{ft/sec}^2) = 1\ \text{poundal}$$

C. **The British Engineering System**

Whereas the primary quantities of the British absolute system are length, time, and *mass*, those of the British engineering system are length, time, and *force*.

Primary Unit

Pound (force): Unit of force. It is the force with which a 1-lb mass (as defined in *B*) is attracted toward the center of the earth at a locality where the acceleration of gravity is $g = 32.174$ ft/sec^2; or in other words, **the pound force is the weight of the standard mass of 1 lb approximately at the latitude of 45°.**

Derived Unit

Slug: Unit of mass.

Newton's second law: $m = \frac{F}{a}$

The unit mass of 1 slug is that mass to which a unit force of 1 lb imparts the acceleration of 1 ft/sec^2:

$$m = \frac{1 \text{ lb}}{1 \text{ ft/sec}^2} = 1 \text{ slug}$$

A Possible Source of Confusion†

In B a pound is used to designate the unit of *mass.*

In C a pound is used to designate the unit of *force.*

The same word stands for two entirely different things. In order to avoid confusion and, moreover, because the system described in B is rarely used, we shall not use the British absolute system. In future references, the pound will be used to designate a unit of *force* according to system C.

Relation between Mass and Weight

Weight is proportional to mass: $W = gm$. The proportionality constant g is the acceleration of gravity. (We shall use approximate figures for g: 32 ft/sec^2 and 980 cm/sec^2.)

The above relation follows from Newton's second law: Weight is a force, the force exerted upon a body by the earth. We can measure a force exerted upon a body of known mass by determining the acceleration imparted to the body by the force in question: $F = ma$.

The earth at a given location imparts to all bodies irrespective of mass the same acceleration, namely, approximately

$$a = g = 32 \text{ ft/sec}^2 = 980 \text{ cm/sec}^2; \qquad \text{hence, weight } W = F = mg$$

The following examples show how the mass of a body can be computed if its weight is given and vice versa:

Cgs system: You are given a 5-gm body. What is its weight? $W = 5 \times 980 = 4{,}900$ dynes.

---

† We shall make occasional use of the kilogram as a unit of force. The kilogram force is defined as the force which imparts to the mass of 1 kg the acceleration of $g = 980$ cm/sec^2. It is the force of gravity acting on the mass of 1 kg at a specified location.

B.E. system: You are given a 64-lb body. What is its mass? $M = W/g = 64/32 = 2$ slugs.

APPENDIX III

Physical Dimensions

All concepts of mechanics can be defined in terms of the basic undefined concepts: mass, length, and time. For example, velocity is defined as distance over time. We call the term which expresses the definition of the derived concept in terms of the basic undefined concepts its physical **dimension.** Thus, the dimension of velocity is distance/time. It is customary to use capital letters M, L, T for mass, length, and time in expressing dimensions of physical quantities. The dimension of velocity is thus written L/T.

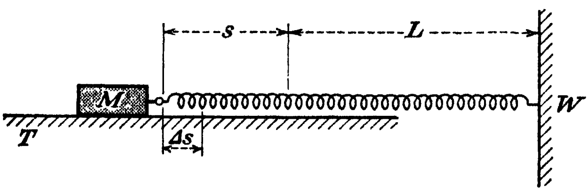
To cite another example, acceleration is defined as change of velocity per unit time. Its dimension is accordingly expressed by the dimension of velocity divided by the dimension of time: $\frac{L/T}{T} = \frac{L}{T^2}$.

We can now obtain the dimension of force by expressing force in terms of mass and acceleration: $F = ma$. The dimension of force must be equal to mass times the dimension of acceleration or $M(L/T^2)$.

Dimensional considerations are a very useful means of establishing the correctness of physical equations. A physical equation cannot be correct unless the quantities which are added or subtracted in the equation have the same physical significance, *i.e.*, the same dimension. This implies that in an equation of the form $A = B$, A and B must have the same dimension.

Example 1

A 16-lb body (M) is sliding on a perfectly smooth horizontal table T. A spring is attached to the body M and to the wall W. The length of the spring is increased $s = 6$ in. over the original length of $L = 2$ ft of the spring. The spring constant is $k = 2$ lb/in. When the body is released for $t = 0.01$ sec, it moves a very short distance Δs which we can consider negligible as compared with s. How large is Δs?



Solution. We are tempted to argue as follows: The mass M is accelerated by a force F. The mass is accelerated in accordance with Newton's second law:

$$F = Ma \qquad (a)$$

The force which accelerates M is due to the distention of the spring and is given by Hooke's law:

$$F = ks \tag{b}$$

Since Δs is negligible as compared with s, we assume s (and hence also F) to remain constant throughout the interval of 0.01 sec during which we consider the motion of M.

Substituting the value of F from equation (b) in equation (a) we obtain

$$Ma = ks \qquad \text{or} \qquad a = \frac{ks}{M} \tag{c}$$

Now for uniformly accelerated motion the distance Δs traversed in time t is

$$\Delta s = \tfrac{1}{2}at^2 \tag{d}$$

Substituting the value of a from (c) in equation (d), we get

$$\Delta s = \frac{1}{2}\left(\frac{ks}{M}\right)t^2 \tag{e}$$

This equation expresses the quantity for which we are looking in terms of given data, and hence the problem is solved.

For obtaining the numerical solution we must be careful to use correct units. Since we are using the British Engineering System, forces and weights are to be measured in pounds, masses in slugs, and distances in feet.

k is given as 2 lb/in., and s as 6 in. The product

$$ks = 2\ \text{lb/in.} \times 6\ \text{in.} = 2 \times 6\ \text{lb/in.} \times \text{in.} = 12\ \text{lb}$$

If we convert these data from inches to feet, we get

$$k = 2\,\frac{\text{lb}}{\text{ft}/12} = 2 \times 12 \times \frac{\text{lb}}{\text{ft}} \qquad \text{and} \qquad s = 6\ \text{in.} = \frac{1}{2}\ \text{ft}$$

hence, $ks = 2 \times 12 \times \text{lb/ft} \times \frac{1}{2}\ \text{ft} = 12\ \text{lb}$, or the same value as above. In this case the conversion from inches to feet was superfluous because the unit of length (feet and inches, respectively) occurs in the denominator of the unit in which k is measured (pounds per foot or pounds per inch), so that the units in the denominator cancel against the units of s by which k is multiplied.

We are told in the problem that M is a 16-lb body. This is obviously the weight of the body. We obtain its mass from the relation

$$m = \frac{W}{g} = \frac{16}{32} = 0.5\ \text{slug}$$

Now we are ready to substitute numerical values in equation (e):

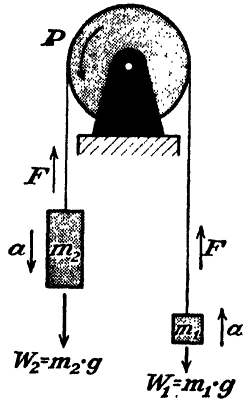
$$\underline{\Delta s} = \frac{1}{2} \times \left(\frac{2 \times 6}{0.5}\right) \times (0.01)^2 = \underline{0.0012\ \text{ft}}$$

Critique. Our argument in which we considered $F = ma$ as a law of motion according to which the mass M has to move was naïve but perfectly acceptable and convenient. Actually, as we remember, $F = ma$ is the definition of F in terms of m and a. We could have solved the problem arguing as follows: The force exerted upon M is given by Hooke's law

$$F = ks \tag{b}$$

which is an empirically found law of nature. We express F in terms of the measurable quantities m and a from the definition equation (a): $F = ma$. We thus get equation (c): $Ma = ks$. From here on the procedure is exactly the same as before.

In general, Newton's second law is used as a definition equation which enables one to replace F by the expression ma in the expression of the force law, such as Hooke's law, or the electrical law of Coulomb, which applies to the problem in question.



Example 2. Atwood's Machine

The masses $m_1 = 48$ gm and $m_2 = 50$ gm are attached by means of a practically massless string to a frictionless pulley P whose mass can be considered as practically zero. The weights are accelerated as shown. Compute the acceleration a.

Solution. Consider mass m_1. It is accelerated upward despite gravity. Where does the accelerating upward force come from? The string attached to m_1 is slightly distended and hence exerts an upward force upon it. The unbalanced force upon m_1 is $F - W_1 = (F - m_1g)$. According to Newton's second law, *i.e.*, Newton's definition of unbalanced forces, this expression is equal to the mass m_1 times the acceleration a of this body:

$$(F - m_1g) = m_1a \tag{1}$$

Similarly we can express the unbalanced force acting on m_2, which is composed of its weight $W_2 = m_2g$ and the upward pull F of the string, as the product of mass times acceleration:

$$(m_2g - F) = m_2a \tag{2}$$

We can now eliminate F from equations (1) and (2) and thus obtain a single equation for a.

From (2) follows $F = m_2g - m_2a = m_2(g - a)$. Substituting this value of F in equation (1), we get

$$m_2(g - a) - m_1g = m_1a$$

or

$$m_2g - m_2a - m_1g = m_1a$$

This is the same as

$$a(m_2 + m_1) = g(m_2 - m_1)$$

from this follows:

$$\underline{a} = g\frac{m_2 - m_1}{m_2 + m_1} = 980\frac{50 - 48}{50 + 48} = 980\frac{2}{98} = \underline{20 \text{ cm/sec}^2}$$

This is the value of the acceleration one would predict for the given masses.

A misconception to be avoided: Suppose we perform the above experiment and find experimentally a value for a which agrees with the predicted value within the limits of experimental error. *Does it mean that we have proved Newton's second law?* As we know, this question is absurd, since Newton's "second law" is a definition and hence incapable of proof. *What we have proved is the fact that the force of gravity on a mass m is given at the locality in question by the product mg, where g = 980 cm/sec².* As we shall see after reading Chap. 13, the equation $W = m \times 980$ is only an approximation. Strictly speaking, the weight of the masses m_1 and m_2 varies (owing to a change in g)

during the experiment as their distance from the center of the earth changes in the course of their motion. But this variation in the value of g with the location of the mass above ground is too small to be detected in a crude laboratory experiment of this kind.

Thus we see that the Atwood machine, the principle of which has been illustrated in this problem, is essentially a device for measuring the acceleration of gravity g by the determination of the acceleration a rather than a setup for the verification of Newton's second law.

QUESTIONS AND PROBLEMS

1. A fisherman has a 2-kg fish on his fishing hook. How large is the tension in the fishing line when (*a*) the fish is suspended on the line at rest? (*b*) the fish is pulled upward with the constant velocity of 9.8 cm/sec? (*c*) the fish is allowed to descend with the constant speed of 980 cm/sec?

2. A 1-kg body which can slide on a frictionless surface is accelerated at the rate of 1 ft/sec^2. Find the accelerating force (*a*) in pounds and (*b*) in dynes.

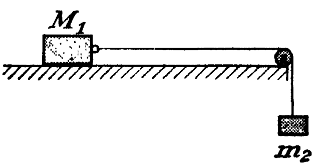
3. A 3,200-lb automobile moving initially at the rate of 30 ft/sec is stopped 3 sec after the brakes are applied. How large is the retarding force?

4. A 6,400-lb truck moving initially at the rate of 40 ft/sec is stopped 60 ft beyond the point at which the brakes are applied.

a. How long does it move while retarding?

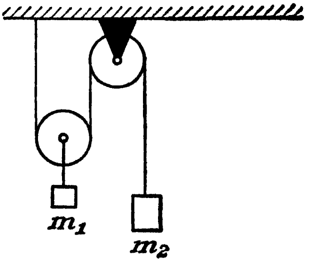
b. How large is the retarding force?

5. The mass $M_1 = 9$ kg rests on a frictionless table top. The mass $m_2 = 1$ kg descends owing to the pull of gravity. Find the acceleration of M_1. (Assume the pulley to be massless and frictionless.)



6. A rope whose breaking strength is 22 lb is to be used to pull a pail of water from a well, lifting it with an upward acceleration of 0.5 ft/sec^2. If the pail plus the water weighs 20 lb, will the rope break?

7. How large a horizontal force should be applied to a 140-lb man on a 20-lb bicycle in order to impart to him the speed of 12 ft/sec within 10 sec? How large a retarding force should we apply in order to reduce his top speed of 12 ft/sec to 6 ft/sec while he traverses the distance of 96 ft? (Neglect friction.)



★8. The masses $m_1 = 1$ kg and $m_2 = 4$ kg are suspended with the aid of string and pulleys as shown in the figure. Neglecting friction and the mass of the pulleys, find the velocities of m_1 and m_2 at the end of the first second of their motion.

★9. In Example 2 we assumed that the tension in the string is the same on both sides of the pulley. This assumption is strictly correct only if the mass of the string is zero. How could you justify the assumption of uniform tension in the string on the basis of Newton's second law?

10. In a slingshot the force accelerating the stone diminishes as the rubber band contracts. We can imagine the final velocity to have been imparted to the stone by a fictitious constant average force acting on the stone while the rubber band contracts. Assuming a rubber band originally 10 cm long distended to twice its original length and a stone of 10 gm mass, find the average force which would impart to the stone the final speed of 10 m/sec.

CHAPTER 5

DYNAMICS: 2. NEWTON'S THIRD LAW AND CONSERVATION OF MOMENTUM

a. Can a body exert a force upon two different bodies at the same time?

b. Does the existence of forces presuppose the existence of material bodies?

c. Can a body A exert a force of 10 lb upon a body B while B exerts only (1) 5 lb, (2) 0 lb upon body A?

d. When you throw a stone, thereby accelerating it forward, how many forces are acting at the point of contact between hand and stone?

e. If forces always occur in pairs of equal and opposite forces, how do we ever manage to accelerate a body?

f. Does every falling object eventually reach a constant terminal velocity?

g. When an apple is released to fall toward the surface of the earth, where does the earth meet the apple?

h. Is the term "impulse" applicable to the product of force times time only if the time interval is very short?

i. Is the law impulse = change of momentum ($Ft = mv_2 - mv_1$) a fundamental law which cannot be derived from Newton's laws of motion?

j. From what fundamental laws can you deduce the law of conservation of momentum?

k. How would you describe the deformation of two colliding bodies in case they are (1) perfectly elastic, (2) perfectly inelastic, (3) imperfectly elastic?

l. What is the distinction between elastic and inelastic impact?

m. Why can't we lift ourselves by our own bootstraps?

n. What is the advantage of conservation principles from the standpoint of economy of thinking?

o. Does the law of conservation of momentum hold equally for inelastic as well as elastic impact?

What Is a Force?

Previously we had asked the question: "What is a force?" and answered it by defining a force as a cause of acceleration.

Is there another way of answering that question which might bring out some other important characteristic of the force concept?

A new approach to the above question is suggested by asking the following: If there were a single mass point A in the universe, could it be acted upon by a force? Obviously, it could not. It would take at least one other body B to exert this force upon it. Thus, a force appears to be an expression for a mechanical interaction between (at least) two bodies. Is this interaction symmetrical, or can body B exert a force upon body A without itself being acted on by A? And if the interaction is always mutual, how do the forces F_A (exerted upon A) and F_B (exerted upon B) depend on the properties of the bodies A and B and their disposition in space and on the possible presence of a third body C?

Newton's Third Law of Motion

In answer to this question Newton enunciated a fundamental law:

If a body *A* exerts a force upon body *B*, body *B* reacts by exerting an equal but opposite force upon body *A*. (To every action there is an equal opposite reaction.)

A noteworthy feature of this law is the implication that **a single, isolated force does not exist in nature.** Forces always occur in pairs of equal but opposite forces (**action** and **reaction**). The number of forces in the universe is even.

Newton's Third Law in Statics

This third law of Newton appears quite plausible when we limit our consideration to bodies at rest. For instance, a book on the table is attracted to the center of the earth and might be expected to fall with acceleration g. The fact that it does not do so implies that the unbalanced force acting upon it is zero; *i.e.*, there must be an upward force acting upon it equal to its own weight in magnitude but of opposite direction, thus balancing the weight in its tendency to produce a downward acceleration. This force is evidently exerted upon the book by the only body in contact with it: the table. The table is slightly deformed under the load of the book until the upward elastic counterforce of the table equals the weight of the book. Then the process of deformation stops, and the book remains at rest. (Actually the table will vibrate slightly before coming to rest.) Thus, the book is at rest with two forces acting upon it: the gravitational pull of the earth and the reaction of the table. The force exerted by the earth pulling the book toward the earth has as its equal and opposite reaction a force exerted by the book upon the earth pulling the earth toward the book. The force exerted by the book on the table has as its equal and opposite reaction the force exerted by the table on the book. The force which the table exerts upon the book and the force of the book on the table (both equal to W) constitute a pair of **action** and **reaction.**

Equilibrium of a body does not imply that the body is necessarily at rest but rather that all forces acting upon it cancel each other's effect so that the body, if not at rest, is in uniform linear motion.

Fall in a Resistive Medium

Parachutes or small bodies such as raindrops acquire a constant velocity soon after release. This means that the unbalanced force acting on them becomes zero. One of the acting forces is that of gravity, $F = \downarrow W$. To balance this force the air must exert upon the raindrops (or parachute) an upward force R equal but opposite to their weight. $R\uparrow = W$ (air resistance). In our first example the book was exerting a force equal to its weight upon the earth through the "medium" of the table. In the present case the atmosphere can be considered as part of the earth. The raindrop presses upon the earth's surface through the medium of the atmosphere, exerting a force equal to its own weight (the force of air friction being transmitted to the surface of the earth). The earth, in turn (the atmosphere being part of it), exerts a reaction (air resistance) upon the drop equal to its weight, which is evidenced by the constant **terminal velocity** reached by the drop. The raindrop is in equilibrium under the action of two forces: the gravitational force of the earth on the drop, which pulls the drop toward the earth (this force has as its equal and opposite reaction the force exerted by the drop upon the earth pulling the earth toward the drop), and the resistance force of the air exerted on the drop pushing the drop upward (this latter force has as its equal and opposite reaction the force exerted by the drop on the air pushing the air down).

All objects falling through a fluid medium (such as air or water) eventually acquire a constant terminal velocity owing to friction, which increases with the velocity. Small objects (having a large cross section as compared with their volume) reach their terminal velocity v sooner (*i.e.*, at a lower value of v) than larger bodies of the same material. In the case of the parachute a large cross section is artificially created by its familiar design in order to diminish the terminal velocity. (We shall consider in greater detail oil droplets falling with their terminal velocity in connection with Millikan's method of measuring the electronic charge.)

Biological and Geological Examples. Certain animals and the seeds of certain plants utilize the parachute principle in order to fall slowly to the ground.

Sedimentation of small dense particles in water is determined by their terminal velocity. The rate of settling is affected not only by size but also by the density of the material: For equal particle size the more dense material will settle down faster. This effect is utilized in placer mining to separate light sand grains from the heavier gold deposit. Medical diagnosis pays attention to sedimentation of blood corpuscles in vitro by measuring their "sedimentation rate." (One of the methods determines the number of millimeters descended by the demarcation line of the blood corpuscle suspension during the first hour.) Except for pregnancy (when the sedimentation rate may increase from 15 to about 45 mm/hr) an increased sedimentation rate usually indicates infection or malignancy. The sedimentation rate, among other factors, is determined by the curious tendency of red blood corpuscles to stick to each other forming the so-called *rouleaux* aggregates.

Newton's Third Law in Dynamics

Whereas the meaning of Newton's third law appears quite plausible as applied to statics (unaccelerated bodies), in application to dynamics it looks at first startling: If you accelerate a stone throwing it eastward, you obviously are exerting an eastward force upon it. But according to Newton's third law this action should bring into play a westward force of reaction of equal magnitude! Should not these two forces, action and reaction, cancel each other, thus making it impossible to throw the stone?

This confusion will be resolved by an important supplementary statement: **Action and reaction do not act on the same body;** *i.e.*, if the action is a force exerted by body A on body B, the reaction will be exerted by B upon A. **(Two forces may cancel each other only if they act upon the same body!)** In our example action is the force our hand exerts upon the stone and the reaction is the backward force, which we can feel the stone exert upon our hand as we throw it.

As another example of action and reaction we consider a heavy apple. If the apple weighs 1 lb, it not only means that the earth pulls it toward its center with a force of 1 lb but it equally means that the apple pulls the earth with the force of 1 lb toward its own center. The earth is very light indeed on the surface of an apple!

If we release the apple, it starts falling toward the earth and the earth starts falling toward the apple. How can we prove that the force exerted by the large earth upon the tiny apple is the same as the one exerted by the small apple upon the enormous earth? In other words, what evidence do we have for the general validity of Newton's third law in dynamics?

Inference of Newton's Third Law from Experimental Evidence. We can infer its validity from the following considerations:

Preliminary Considerations. Imagine a force F applied to a mass m for such a short time t that F can be considered as constant during this interval.

Newton's second law: $$F = ma = m\frac{v_2 - v_1}{t} \tag{5.1}$$

multiply by t:

$$\boxed{\overrightarrow{Ft} = \overrightarrow{mv_2} - \overrightarrow{mv_1}} = \overrightarrow{\Delta mv} \qquad (5.2)\dagger$$

(Impulse = final momentum − initial momentum)

---

† The horizontal arrows symbolize that this is a vector equation. The student may disregard this feature in the first reading but should reconsider the meaning of this equation after studying Chap. 7.

This is an important expression. It is apparent from its derivation from Newton's second law that it makes no essentially new statement. It implies that, when you apply a constant force F for a time t to a mass m, you increase its velocity from v_1 to v_2. (The product of force times time has received a special name, **impulse,** and the product of mass and velocity, **momentum.** Note that impulse and momentum are both vector quantities.) Equation (5.2) states: **The change of momentum of a body is equal to the impulse which caused it.**

Derivation. Let us now imagine two masses m_1 and m_2 pulled toward each other (or pushed apart) by a (massless) spring between them (Fig. 5.1).

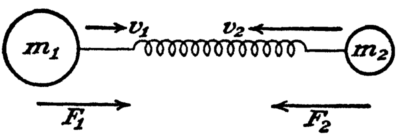


FIG. 5.1. Masses pulled toward each other by a spring.

The force acts for the same time t on both bodies. But we don't know whether or not F_1 is equal to F_2. If the spring is allowed to act for only a short time t (so that F remains practically constant) and is then moved out of the way, the bodies which started from rest will have acquired the final velocities v_1 and v_2. We can apply equation (5.2) to each of these bodies:

$$F_1t = m_1(v_1 - 0) \tag{5.3a}$$
$$F_2t = m_2(v_2 - 0) \tag{5.3b}$$

Dividing (5.3*a*) by (5.3*b*) we get

$$\frac{F_1t}{F_2t} = \frac{m_1v_1}{m_2v_2} \tag{5.4}$$

Now experience shows that, in all kinds of experiments on interaction between two bodies, the momenta acquired by the bodies are equal to each other in magnitude and opposite in direction:

$$m_1v_1 = -m_2v_2 \tag{5.5}$$

or

$$\frac{m_1v_1}{m_2v_2} = -1 \tag{5.6}$$

Substituting equation (5.6) in (5.4) we get

$$\frac{F_1}{F_2} = -1 \tag{5.7}$$

or

$$\boxed{F_1 = -F_2} \tag{5.7a}$$

which expresses **Newton's third law.**

Thus, the validity of Newton's third law has been inferred from the fact that the **sum of the momenta which two interacting bodies impart to each other is always zero.** This statement is merely a different way of expressing equation (5.5), which could also have been written

$$m_1v_1 + m_2v_2 = 0 \tag{5.5a}$$

Law of Conservation of Momentum

Equation (5.5*a*) is a statement of the so-called *law of conservation of momentum* for a special situation, namely, for the interaction of two masses m_1 and m_2, initially at rest, which acquire the velocities v_1 and v_2 as a result of their interaction. This law is, however, universally valid regardless of the initial velocities of the two interacting bodies, which may be moving in different directions during the interaction. **The law of conservation of momentum asserts that, whenever one of the two interacting bodies gains a momentum $\Delta(\overrightarrow{mv})_1$, the other body gains an equal and opposite momentum $\Delta(\overleftarrow{mv})_2$ so that $\Delta(\overrightarrow{mv})_1 + \Delta(\overleftarrow{mv})_2 = 0$.**[†] **Thus the interaction cannot change the original momentum of the system.**[‡] This is expressed by saying that **the momentum of the system is conserved (unchanged) in the absence of external forces,** *i.e.*, forces exerted by a third body which is not a part of the system (*i.e.*, group of bodies) considered.

Experience shows that **the law of conservation of momentum applies to systems composed of any number of bodies.** This means that the momentum of the system can be changed only by forces exerted upon bodies composing the system by bodies located outside the system.

("Internal forces," *i.e.* forces of interaction between bodies of a given system, such as the forces occurring in collision, do not affect the momentum of the system, since they occur in pairs of action and reaction which are equal and opposite.)

The law of conservation of momentum is of fundamental importance in mechanics. It will enable us, among other things, to solve impact problems, which play an important role in the study of atomic physics and of kinetic molecular theory.

As an illustration of the law of conservation of momentum, we shall consider a case of collision.

[†] Vectors pointing in opposite directions being taken with the positive and negative sign, respectively.

[‡] That is, the vector sum of the original momenta of the interacting bodies (see Chap. 7).

Let us imagine two billiard balls colliding as shown in Fig. 5.2. During the period of contact the balls are slightly flattened and move with the same velocity v. As they straighten out to assume their original shape, the two balls are pushed apart by equal and opposite forces, which change their original velocities from V_1 to V_2 and from v_1 to v_2, respectively. An impulse of the same magnitude Ft changes the momenta of the two bodies. (To avoid the use of calculus we assume that a constant average force F acts during the impact.) We can apply equation (5.2) to each of the two masses:

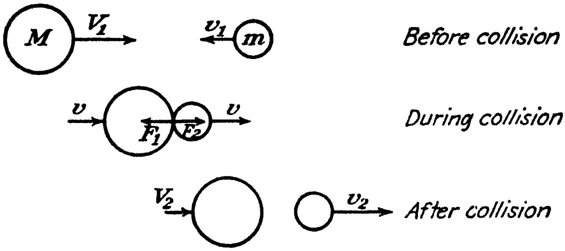


FIG. 5.2. Elastic collision between two spheres.

$$\overrightarrow{F_1t} = \overrightarrow{MV_2} - \overrightarrow{MV_1} \qquad (5.8a)$$

$$\overrightarrow{F_2t} = \overrightarrow{mv_2} - \overrightarrow{mv_1} \qquad (5.8b)$$

Since $F_1 = -F_2$, $\overrightarrow{MV_2} - \overrightarrow{MV_1} = -(\overrightarrow{mv_2} - \overrightarrow{mv_1}) = \overrightarrow{mv_1} - \overrightarrow{mv_2}$ (5.9)

Rearranging terms, $\boxed{\overrightarrow{MV_1} + \overrightarrow{mv_1} = \overrightarrow{MV_2} + \overrightarrow{mv_2}}$ (5.10)

This is a statement of the law of conservation of momentum. The process of impact has not changed the sum of the momenta of the interacting bodies. It was $\overrightarrow{MV_1} + \overrightarrow{mv_1}$ before the impact, and it is $\overrightarrow{MV_2} + \overrightarrow{mv_2}$ after the impact. Our conservation equation states that these two sums are equal.† If we know the masses M and m, their initial velocities V_1 and v_1, and the final velocity of one of them, say V_2, we can determine from equation (5.10) the unknown remaining final velocity v_2 of m.

Conservation Principles in Physics

The law of conservation of momentum is the first conservation principle (*i.e.*, a principle which asserts that a certain parameter of the system remains constant) we have encountered in physics. We shall meet later other principles of this kind, such as the laws of conservation of energy and conservation of mass. Such general conservation laws are very powerful scientific tools. They permit the explanation of facts and solution of problems without considering the details of a process. In fact, often we can determine the result to be expected without understanding the process by which it comes about.

---

† Problems of this type will be treated more extensively in connection with the discussion of the law of conservation of energy (Chap. 10).

Example of the Use of a Conservation Principle. For example, a hunter shoots a bird. The mass of the fowl is 1 kg; that of the bullet is 10 gm. The bird flies east at 10 m/sec, while the bullet travels east at 500 m/sec. What is the eastward velocity of the bird after death?

What Do We Know? Initial mass, eastward velocity, and, hence, eastward momentum of the bird and of the bullet. The bullet remains embedded in the bird. The bird suffers an eastward acceleration as a result of the impact of the bullet. The bullet is decelerated.

What Don't We Know? The common eastward velocity of the dead bird with the embedded bullet, the accelerating force the bullet exerted upon the bird (and, hence, also the decelerating force exerted by the bird upon the bullet), the time of interaction between the bird and the bullet.

The task of observing the time and force and mode of interaction between the bird and the bullet looks hopelessly difficult. There are three unknowns, and we should have to set up three equations to solve this problem. The law of conservation of momentum enables us, however, to solve this problem using only one equation without considering the details of the collision process. We merely have to know the total eastward momentum of the system (consisting of bullet and bird) before the collision and remember that it persists unaltered after the collision:

$$\begin{array}{ccccccc} MV_1 & + & mv_1 & = & (M+m) & & [V] \\ \begin{array}{c}\text{Initial}\\ \text{momentum}\\ \text{of bird}\end{array} & + & \begin{array}{c}\text{initial}\\ \text{momentum}\\ \text{of bullet}\end{array} & = & \begin{array}{c}\text{mass of}\\ \text{bird and}\\ \text{bullet}\end{array} & \times & \begin{array}{c}\text{common}\\ \text{final}\\ \text{velocity}\end{array} \end{array}$$

The only unknown in the above equation is V, and thus our problem is solved. Substituting the given values into the above equation using grams and meters per second, we obtain

$$1{,}000 \times 10 + 10 \times 500 = (1{,}000 + 10)V$$

and
$$V = \frac{10{,}000 + 5{,}000}{1{,}010} = \frac{15{,}000}{1{,}010} \cong \underline{14.85\ \text{m/sec}}$$

for the speed of the shot bird.

Classification of Impacts. The above example illustrates the so-called **perfectly inelastic impact,** *i.e.*, a collision after which the two bodies move together with the same common velocity.

Another extreme case is the so-called **perfectly elastic impact.** It is characterized by the fact that the relative velocity between the two bodies A and B is the same in magnitude after the collision as it was before the collision but opposite in direction, so that two bodies which approach each other before collision recede from each other after collision. Impacts between atoms and molecules are usually (though not always) of this type. In ordinary practice we commonly encounter **imperfectly elastic impact,** in which case the relative velocity is diminished after the collision.

Illustrations and Applications. The following phenomena are explained by the fact that, when body A ejects body B, it will acquire an additional momentum equal to the momentum imparted to B but opposite in direction (conservation of momentum of the system consisting of A and B).

1. *Recoil of a Gun.* The original momentum of the system (gun + bullet) is zero. For the final momentum after the shot to remain equal to zero, the gun must acquire a backward momentum equal in magnitude to the forward momentum of the bullet.

2. *A rocket* utilizes the recoil principle, but instead of bullets it ejects a steady stream of gas. Figure 5.3 shows the diagram of a V-2 rocket used by the Germans during World War II. The continuous stream of the gaseous combustion products of alcohol is expelled through an opening in the back of the rocket. The steady stream of escaping gas imparts a continuous thrust to the rocket, which is thus accelerated. This rocket was designed to move through the stratosphere where the air resistance is low and the oxygen supply is very scarce. Therefore it carried its own supply of liquid oxygen. Similar rockets may gain importance in peacetime as rapid means of communication.

3. *An airplane or a bird or a fly* utilizes the recoil principle to keep aloft. They must throw downward with their wings a stream of air to acquire an upward impulse to overcome the force of gravity. (Details like air circulation about the wings will be discussed in the chapter on fluid mechanics.)

4. *The "ballistocardiograph"* makes use of the recoil suffered by the entire human body when a spurt of blood is ejected by the heart into the aorta during the systole. You can notice this effect when you weigh yourself standing on the platform of a spring scale. The pointer is seen distinctly to oscillate in the rhythm of the heartbeat. A satisfactory mathematical analysis of this phenomenon has not yet been carried out because of numerous complications, but an empirical correlation between the observed mechanical oscillations and the cardiac output (the rate at which the heart pumps blood) is possible and promises to be of clinical value.

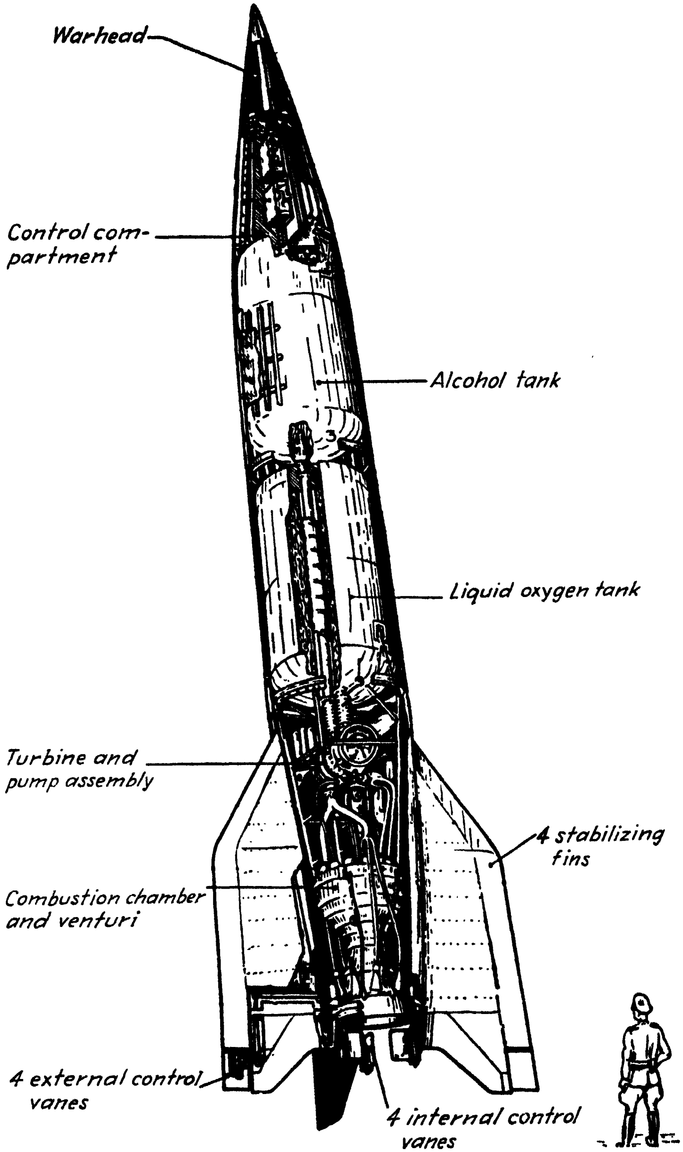


FIG. 5.3. Sectional view of a V-2 rocket used by Germany for long-range bombardment in World War II. The explosive is located in the "warhead." Alcohol serves as fuel. Liquid oxygen makes combustion possible at great altitude (where the air is rarefied). Both liquids are supplied to the combustion chamber by turbine-driven pumps. The hot gases resulting from the combustion of alcohol escape at the lower end imparting to the rocket a thrust of some 70,000 lb by which the rocket is propelled. (*From Kooy and Uytenbogaart, "Ballistics of the Future," McGraw-Hill.*)

Reconsideration of Newton's Laws

Let us now reconsider Newton's second and third laws in order to see whether or not they are consistent with each other.

Suppose we have two men 1,000 yd apart whose masses are m_1 and m_2. Assume that they are the only two bodies in this universe. The two men are joined by a rope. The first man pulls the rope, thus exerting a force F upon the second man and experiencing himself an equal force of attraction toward the other man according to Newton's third law.

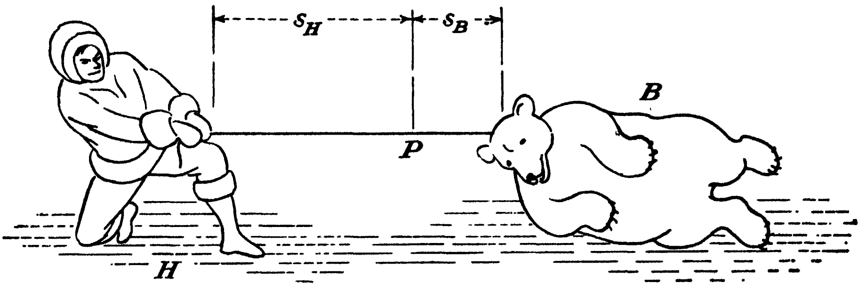
In order to check the validity of Newton's laws each of the observers determines his acceleration, multiplies it by his mass, and checks whether or not $m_1a_1 = m_2a_2$.

The only two bodies in the universe to which the observers can refer their acceleration are m_1 and m_2, *i.e.*, the observers themselves. Suppose the first observer argues as follows: I assume arbitrarily that I remain at rest (my velocity and acceleration remaining zero). I can then determine the acceleration a_2 of the second man with respect to me by measuring the change of our separation in the course of time. The product of mass times acceleration is then: for me, $F_1 = m_1a_1 = m_10 = 0$ and, for my colleague, $F_2 = m_2a_2 \neq 0$; hence, $F_1 \neq F_2$. Using the other man as a reference point of zero acceleration leads to the same result. Something seems to be wrong with one or both of the laws we used or with our formulation of the laws.

It is easy to see that the difficulty lies in the vagueness of the reference to acceleration in our previous formulation of Newton's second law: *Force is equal to mass times acceleration.* Acceleration with respect to what? A body which remains forever at rest (or in a state of uniform motion) would be our ideal reference body of zero acceleration. But where can we find such an unaccelerable body? Any sufficiently large mass will do for this purpose. The acceleration imparted by a force to a body is inversely proportional to the mass. The acceleration due to a finite force would be zero for an infinite mass. The mass of the earth would be large enough to be considered as infinite in comparision with any terrestrial body. Even better would be the mass of the entire universe of stars, which can be considered practically as infinitely larger than the mass of any individual planet or star. We can choose the center of this huge mass as our unaccelerable reference point and *amend Newton's second law by the statement*: *a* **is the acceleration with respect to the center of mass of the universe.** When so amended Newton's law gives a consistent description of our experience. *In practice it is sufficiently accurate for most purposes to determine the acceleration of bodies* *relative to the earth's surface.*† In the above example of a starless universe the inconsistency is removed when the mass center‡ of m_1 and m_2 is used as reference point.

Example

An Arctic explorer tries to pull a dead polar bear toward himself by means of a rope. Both he and the bear are on perfectly smooth ice. The explorer weighs 150 lb and is 120 ft away from the bear. He gives the rope a quick pull (we shall interpret this to mean that he exerts a constant force for a brief time interval) so that he and the bear start sliding toward each other. The hunter and the bear collide 30 ft from the original position of the bear. How heavy is the bear?



Solution. The impulse exerted upon the bear is the same as the impulse exerted upon the hunter, according to Newton's third law. For the hunter:

$$Ft = m_H v_H \tag{1}$$

where m_H is the mass of hunter and v_H the final velocity he acquires as a result of the impulse. Similarly we can write for the bear

$$Ft = m_B v_B \tag{2}$$

Equating (1) and (2) we obtain

$$m_H v_H = m_B v_B \tag{3}$$

or

$$m_B = m_H \frac{v_H}{v_B} \tag{3a}$$

Since the hunter and the bear start and stop moving at the same time, the time of motion T is the same for both. The distance the hunter moves from his starting point to the collision point P is $S_H = v_H T$, and for the displacement of the bear we can write $S_B = v_B T$.

Let us now multiply the right side of equation (3a) by $T/T = 1$. We get

$$m_B = m_H \frac{v_H T}{v_B T} = m_H \frac{S_H}{S_B} \tag{4}$$

---

† The surface of the earth is not strictly unaccelerated. We shall see in Chap. 7 that points on a rotating body possess a "centripetal" acceleration. The acceleration on the surface of the rotating earth is small enough to be neglected in most practical cases.

‡ That is, the point at which the two interacting masses collide eventually. This point divides the distance between the two masses in the inverse ratio of the masses (see example).

Now our problem is solved! $S_B = 30$ ft is given in the problem. S_H follows from the fact that the original distance between the bear and the hunter is 120 ft $= S_H + S_B$.

Hence, $$S_H = 120 - 30 = 90 \text{ ft}$$

Substituting the given data in (4) and multiplying both sides by the acceleration of gravity g we get

$$\text{Weight of the bear} = \underline{(m_B g)} = (m_H g)\tfrac{90}{30} = 150 \times \tfrac{90}{30} = \underline{450 \text{ lb}}$$

We note from equation (4), which can be written $S_H/S_B = m_B/m_H$, that the collision point P divides the original distance between the two bodies in the inverse ratio of their masses. The point P is called the *center of mass* of the two-body system.

QUESTIONS AND PROBLEMS

1. With what velocity would a raindrop hit our head if it originated in a cloud 400 m high and fell through a vacuum? How does the velocity compare with the speed of a bullet (500 m/sec)? What is the actual motion of a raindrop like as it moves through the atmosphere?

2. Why can't we walk on perfectly smooth ice? On what principle is our way of walking and, in fact, the locomotion of all animals based?

3. What is wrong with this argument: "The law of conservation of momentum is false. For suppose we have two boxcars moving toward each other, one of mass m_1 and velocity v_1 and the other of mass m_2 and velocity v_2. When they collide, they will couple together and move with the same final velocity V. Now everybody knows that V is less than v_1 or v_2 in this case, so that $m_1V < m_1v_1$, $m_2V < m_2v_2$, and

$$m_1V + m_2V < m_1v_1 + m_2v_2$$

Therefore, momentum disappears in the above collision."

4. Can you infer whether or not the force of air resistance exerted upon a falling body is independent of its speed from the fact that the terminal velocity is not reached instantly at the moment of release?

5. Can an elastic ball bounce upward from an elastic floor to a greater height than the one from which it started on its way down?

6. Is it correct to speak of a momentum of 10 lb-sec? (Explain.)

7. A 10-gm stone, initially at rest, is accelerated to a final speed of 30 m/sec by means of a slingshot.

a. What is the momentum imparted to the ball?

b. What is the numerical value of the impulse?

c. What is the value of the average force exerted upon the stone by the slingshot if we assume the force to act for 0.1 sec?

8. A ½-lb ball is given the speed of 60 ft/sec by kicking it with a foot.

a. How large an impulse was imparted to the ball?

b. What is the average value of the force exerted upon the ball if the contact between ball and foot lasts 0.05 sec?

9. A fast luxury liner is equipped with a skating rink. Two passengers suffer a head-on collision while skating. The liner moves east with a speed of 20 ft/sec, passenger A (160 lb) moves east with a speed of 15 ft/sec, and passenger B (200 lb) moves west with a speed of 12 ft/sec (with respect to the ship).

a. Assuming that the passengers cling together after the collision, find their final velocity.

b. What result would you get if the velocity of the ship were twice as high?

10. An anchored armored boat is subjected to a heavy machine-gun fire. Its armor is very good, and all bullets recoil backward like perfectly elastic balls. Assuming the boat to be hit on the average by 100 bullets per second, taking the weight of the bullet as $\frac{1}{20}$ lb and its velocity as 1,200 ft/sec, compute the average force with which the machine gun tends to repel the boat.

11. One atom of helium (mass 4 arbitrary units) collides with an atom of oxygen (mass 16 arbitrary units). If the two atoms are moving originally along the same straight line with the velocities of $v_{\text{He}} = 6 \times 10^8$ cm/sec, and $v_{\text{O}_2} = 2 \times 10^8$ cm/sec, what will be the respective final velocities after an elastic collision of these two atoms?

★12. Two masses m_1 and m_2 collide with each other as a result of mutual attraction. $m_1 = 5\ m_2$. The masses are small enough to be treated as points. If their initial distance is 10 cm, how far from the initial position of m_1 do the masses collide?

13. A small rocket whose mass is 220 gm, including a 20-gm powder charge, is capable of reaching a final velocity of 60 m/sec in horizontal flight. The combustion of the powder charge is completed in 0.2 sec.

a. Estimate approximately the average force which accelerates the rocket (*i.e.*, give limits between which the force must lie).

b. How high would the rocket rise in vertical ascent? (Neglect air resistance.)

14. You are standing at the rear end of a 12-ft-long boat whose weight is 400 lb. The front end of the boat touches the pier. As you walk toward the front, the boat recedes from the pier. The boat stops at a certain distance S from the pier when you reach the front end. Neglecting the resistance of the water, how large is S if your weight is 100 lb? What happens to the location of the center of mass of the boat-passenger system as the passenger changes his position? (Hint: This situation is analogous to the one treated in the example.)

CHAPTER 6

DYNAMICS: 3. SIMILARITY BETWEEN EFFECTS OF GRAVITY AND OF INERTIA

a. How would you describe the feeling that you get in an elevator (1) when it starts accelerating downward? (2) when it retards on its way up? (3) when it retards on its way down? (4) when it accelerates going upward?

b. How would you explain your answers to question *a* mechanically? Which laws of motion are involved?

c. Does the mass of a body appear to change when measured in an accelerated box?

d. Does the weight of a body appear to change when measured in an accelerated box?

e. How would you have to move in order to record zero weight while standing on a spring scale?

f. Would the result be the same as in *e* if you used a balance with weights on the other pan?

g. What happens to your stomach when you jump off a roof?

h. If you put a 10-lb stone on top of a 5-lb stone and drop them simultaneously, how large a force does the top stone exert upon the bottom one?

i. Can a scientist in an accelerated box apply Newton's second law while measuring the acceleration of bodies relative to his box?

j. What is a d'Alembert force?

k. Could you perform experiments on biological effects of high gravitational forces without traveling to a heavier and denser planet?

l. If you swing your watch, holding it by the chain, fast enough, the chain will break. How do you explain it?

m. With what physical phenomenon is aviation medicine confronted in the case of a dive-bomber pilot's "blackout"?

n. How do you explain the centrifugal force which you experience in a merry-go-round?

Example of the Accelerated Elevator. It is a fact of common experience that in an *accelerating* elevator we feel our weight (1) diminish when the elevator (*a*) starts going downward or (*b*) stops going up, (2) increase when it (*a*) starts going up or (*b*) stops going down. (In the first two cases acceleration points *downward.* In the second two cases it points *upward.*†)

† The direction of the acceleration can be found easily from $a = (v_2 - v_1)/t$. If we adopt the convention that a vector $-V$ designates a vector of the same magnitude as V *but of opposite direction,* we can see

Case 1*a*: $v_2 > v_1$; hence, a has the same sign as v_2, *i.e.*, points down.

Case 1*b*: $v_1 > v_2$; hence, a has the same sign as $-v_1$, *i.e.*, has opposite direction to v_1. Since v_1 points upward, a points downward.

Case 2*a*: $v_2 > v_1$; hence, v_2 points up and a points up.

Case 2*b*: $v_1 > v_2$; hence a and $-v_1$ have same direction or the directions of a and v_1 are opposite: v_1 points down, and a points up.

We shall discuss the cases 2*a* and 1*a*, leaving the discussion of the two others to the student.

Case 2a: Elevator Accelerated Upward (Fig. 6.1). When the elevator stands still (or moves uniformly) the *spring-scale* reading is $R = W$, the weight of the load.

The weight rests on the spring-scale platform, exerting upon it the force W and experiencing itself the upward reaction R, which the scale exerts in turn upon the weight. The reading of the scale is equal to the reaction which it exerts. Under the influence of the two equal and opposite forces, W (attraction of the earth) and R (reaction of the scale), the weight remains unaccelerated.

Now the elevator is accelerated upward. We see the weight move with an upward acceleration a. According to Newton's second law this must be due to an unbalanced upward force, which must be equal to the mass of W times its acceleration. This unbalanced upward force F must be the resultant of the only two forces acting on our weight, W and R.

$$\uparrow F = \uparrow R - \downarrow W = ma\uparrow \tag{6.1}$$

Hence,

$$\boxed{R = W + ma} \tag{6.1a}$$

Since R stands for the force of reaction as well as for the reading of the scale, *this means:* The scale reading is increased by ma above normal, that is, *W appears to have gained weight, the gain being equal to ma.*

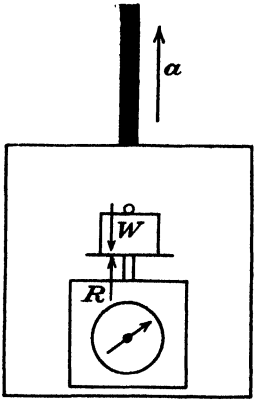


FIG. 6.1. Weighing in an elevator which is accelerating upward.

Case 1a: Elevator Accelerated Downward (Fig. 6.2). Again we fix our attention upon the weight W. It is under the action of two forces only, W and R. There must be an unbalanced downward force acting on it to cause its downward acceleration. According to Newton's second law:

$$\downarrow F = \downarrow W - \uparrow R = ma\downarrow \tag{6.2}$$

Hence,

$$\boxed{R = W - ma} \tag{6.2a}$$

The scale reading R is diminished. Our load W appears to have lost weight! The apparent loss of weight is equal to ma.

Formula (6.2*a*) leads to a striking speculation: If we cut the cable, the elevator will fall freely and a will be equal to g. Then, $R = W - mg$. But as we know [equation (4.3)], $W = mg$; hence, in the special case of

a freely falling box,

$$\underline{R = mg - mg = 0}$$

the scale reads zero, gravity seems to have disappeared. No experiment will reveal to an observer isolated inside a freely falling sealed box the proximity of the gravitating earth. His objects will not fall toward the floor of the elevator, his pendulum won't swing, and all other manifestations of gravity will vanish. The organs within his body will be slightly relocated, because as soon as gravity ceases to act upon his viscera, they will yield to the elastic forces of the tissues from which they are suspended. As a result, he will definitely *feel* that he is falling, but this feeling will be equivalent to the sensation of cessation of gravity.

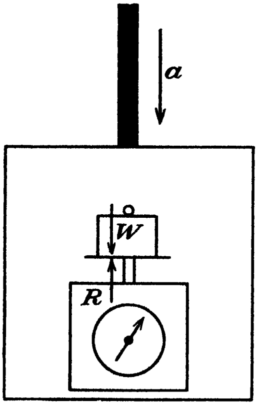


FIG. 6.2. Weighing in an elevator which is accelerating downward.

An observer on firm ground who watches the observer in the falling box weigh a mass m and record a zero weight has no difficulty in explaining these observations in terms of his system of reference. To explain why the weight appears to vanish in the falling box, he argues as follows: W cannot possibly press upon the scale pan. It is falling toward the ground with the same acceleration as the pan. If one were to separate them by 1/1,000 in. they would continue to fall, maintaining this constant separation, and one would continue to have the illusion that the weight "rests" on the scale. The weight could evidently press on the scale only if the latter interfered with its tendency to fall with the acceleration of gravity g.

The observer in the falling box, however, will encounter a difficulty in reconciling his observations with Newton's laws if he uses his falling box as a reference system with respect to which he measures acceleration of bodies. He will notice that a released body does not fall but remains at rest with respect to the box in spite of the acting force of gravity. If he rejects the assumption that gravity ceased to act on the released body, he will be unable to reconcile with Newton's second law the fact that a body remains at rest (with respect to him) despite a force acting on it.

D'Alembert's Principle and Inertial Forces

Here an ingenious idea of d'Alembert can be introduced to maintain the validity of Newton's second law for accelerated reference systems. It will lead us to a new and interesting way of considering processes in accelerated systems, which may enable us in certain cases to answer simply many problems which otherwise would require complex calculations for their solution. D'Alembert introduced **fictitious inertial forces** into the picture (which we shall call **d'Alembert forces**) which restore the validity of Newton's second law for accelerated systems. This purely formalistic procedure leads to a very helpful mental device, since it allows us to convert a problem in dynamics into a usually simpler problem in statics.

Consider the equation $F = ma$. The quantity on the left side of any equation must have the same physical significance (dimension) as the quantity on the right side. Since F is a force, the product ma must have the meaning of a force. The above equation can be also written

$$F - ma = F + (-ma) = 0$$

We shall call $-ma$ the d'Alembert force.[†]

Hence, Newton's second law would read in d'Alembert's terminology as follows: The sum of the force applied to the body and the d'Alembert force is always 0, or the applied force and d'Alembert force are in equilibrium. Thus, if an observer in an accelerated system (box) always adds a (fictitious) inertial force $-ma$ (where a is the acceleration of the box with respect to the ground) to the applied forces, he will find no inconsistencies with Newton's second law if he measures the acceleration of bodies inside his box with respect to the box. In the example of an elevator accelerated *downward*, the force applied to W is its own weight: $W\downarrow$. The d'Alembert force is $-ma\uparrow$. Hence, the sum of the forces observed acting on our body in the accelerated box is $W - ma$, and this is actually the reading recorded by the scale [see equation (6.2a)]. We write in an accelerated box

$$\underset{\substack{\downarrow \\ \text{Unbalanced} \\ \text{force}}}{\underline{(W - ma)}} = \underset{\substack{\downarrow \\ W}}{\underline{mg}} - ma = m\underset{\substack{\downarrow \\ \text{relative} \\ \text{acceleration}}}{\underline{(g - a)}},$$

which is an adaptation of Newton's second law to the accelerated elevator as system of reference.

In the case of the freely falling box the applied force (weight) is W and the d'Alembert force is $-mg$. The resultant total force is $W - mg = 0$. We see from this how the introduction of the d'Alembert force $-mg$ again preserves the validity of Newton's second law with respect to the accelerated observer: The weight is at rest *with respect*

[†] *Note that the direction of the d'Alembert force is opposite to the direction of the acceleration of the box:* If the acceleration is $\overrightarrow{+a}$, the force is $\overleftarrow{-ma}$.

to him, and the addition of $-mg$ makes the unbalanced force equal to 0. Thus

$$F = \underline{W - mg} = ma = \underline{0}$$

[where g is the acceleration of the system (box) relative to the earth and a is the acceleration of the body under consideration (W) relative to the box.]

In order to get a clearer idea of the meaning of d'Alembert forces let us consider the following example (Fig. 6.3): Suppose we have a box in a region free of gravity. All objects will stay put wherever you place them in mid-air. Suppose, now, somebody starts pulling this box west with a constant acceleration a. An outside observer looking through a glass wall into the box will interpret what he sees as follows: The weight W, being attached to the box through the spring S, is accelerated along with the box with the same acceleration a. The force which the spring transmits to W must be equal to ma. The ball B, on the other hand, tends to remain at rest, since no forces can be transmitted to it. (There can be no friction against the floor, since there is no gravity to press the ball against the floor.) Consequently, the east wall of the box approaches B with the acceleration a.

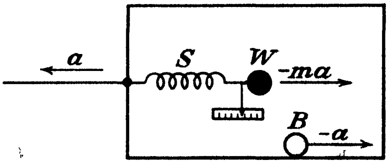


FIG. 6.3. Illusion of a gravitational field in an accelerated box.

The same phenomena will be described quite differently by an observer in the box, who does not know that he is being accelerated. He is likely to assume that the east wall is the floor of his box, which he believes to be standing on the surface of a planet. The spring S is then naturally distended by the weight $W = -ma$, where we recognize $-ma$ as the "d'Alembert weight" of W, the mass of which we call m. He refers all motion to his box, which he assumes to be at rest, and concludes that all objects including the ball B in his box "fall" toward the east wall with the acceleration $-a$. In other words the acceleration of the box gives to an observer in the box the impression of being in a gravitational field.[†] The intensity of attraction may be different from that of the earth and a may not be equal to g, but he has no way of discriminating between the gravitational and accelerational hypothesis by any observations he might perform in his closed box. Unless he looks out of a window and sees what is happening, the assumption that he rests on a planet will be just as plausible to him as the assumption that he is being accelerated in empty space. This equivalence of gravitational and inertial effects plays a fundamental role in Einstein's general theory of relativity.

[†] That is, in a region of gravitational attraction.

Centrifugal D'Alembert Forces

We shall now consider an example of inertial forces which are of great practical importance. (The final convincing argument proving that these forces are of the same kind as the above-mentioned d'Alembert forces will be provided through the discussion of the centripetal acceleration in Chap. 7.)

You know that, if you attach a stone to a string and swing it rapidly in a circle, you may break the string. To succeed in doing this you must either choose the stone heavy enough or swing a light stone fast enough. To break the string the revolving stone obviously has to exert a force upon the string directed away from the center of revolution and stretch it to the breaking point just as a heavy stone stretches a string until it breaks. This stretching of the string involves a slight displacement of the stone away from the center of revolution; we shall call such a displacement a centrifugal displacement (meaning fleeing the center; *fugare*, Latin: to flee).

Why does the string break? Why does the rotating stone behave like a heavy load attached to the string?

We know of two mechanisms which can impart a weight to a body: (1) attraction by another body such as the earth, (2) creation of a d'Alembert force by acceleration.

The first mechanism is obviously out of the question. There is no body pulling the stone away from the center of revolution. But the second mechanism looks at first sight just as hopeless! The assumed d'Alembert force responsible for the breaking of the string is directed away from the center of revolution. Since the direction of an inertial force ($-ma$) is opposite to the direction of the acceleration, we conclude that to cause an outward d'Alembert force we need an inward acceleration (**centripetal acceleration**). Now this is precisely where our difficulty comes in. We are swinging the stone perfectly uniformly. Let us say it makes 2 revolutions per second all the time. How can we speak here of an accelerated motion of the stone? And moreover, how can we visualize an acceleration the direction of which is perpendicular to the motion of the object? The stone moves in a circular arc, and the supposed acceleration must point toward the center!

It will be shown in Chap. 7 that the stone is actually accelerated toward the center of the circle as it moves at uniform speed in a circular arc. We shall anticipate this result and interpret the force which the stone experiences in the direction away from the center of revolution as an inertial force due to the centripetal acceleration of the system similar to the quasi-gravitational d'Alembert force observed in the box of Fig. 6.3.

Biological and Technological Applications. Centrifugal d'Alembert forces have found numerous applications in science, technology, and medicine. One can best understand these applications if one considers that an observer located at some distance from the axis of a rotating body (such as a merry-go-round) will have the impression of being in a gravitational field, the force of gravity being directed away from the center. The quasi-gravitational action of this centrifugal field is demonstrated very strikingly by its effect on plant growth. Plants exhibit as a rule the phenomenon of "geotropism." The stem grows upward, and the roots downward. If a flower pot is inclined so that the stem is horizontal, its tip soon begins to curve and to grow upward, whereas the roots curve so as to grow downward, so that the plant becomes S-shaped. It is, however, possible to suppress this tendency by placing the flower pot on the rim of a merry-go-round, the horizontal stem pointing toward the axis of rotation. If the wheel spins rapidly enough, the stem remains substantially horizontal and grows toward the center of rotation, whereas the roots grow in the opposite direction. The direction toward the axis is now what we would call *upward* and away from it *downward.*

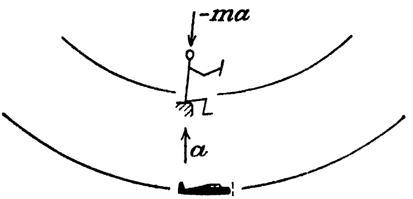


FIG. 6.4. Centrifugal inertial forces acting on a dive-bomber pilot.

What we accomplish essentially by "centrifugation" is to impart to every particle of the rotating body a d'Alembert weight which, in many cases, can be thousands of times as large as the normal gravitational weight of the body. For example, when we swing our arm rapidly about an axis passing through our shoulder, the blood in our hand is pressed strongly toward our fingertips in which the local blood pressure can be made to exceed many times its normal value. Under these conditions a small scratch wound would bleed profusely. Such centrifugal effects upon our blood pressure play an important role in aviation. When a dive bomber is coming out of a dive (Fig. 6.4), the pilot's blood tends to rush toward his feet, reducing temporarily the blood supply to his brain. This renders the pilot temporarily unconscious. (Unconsciousness is preceded by a "blackout" of vision.)

Strong centrifugal d'Alembert forces can also be used to speed up the normally slow rate of sedimentation of small dense particles in suspensions or the rate of accumulation of small light particles, such as oil globules of cream, at the liquid surface. The dense particles which settle at the bottom of the container under the influence of gravity travel toward the peripheral end of the container in the **centrifuge** (Fig. 6.5), whereas the fat globules, which are less dense than water and normally rise to the top of the liquid, travel toward the axis of rotation of the centrifugal cream separator.

A centrifuge such as shown in Fig. 6.5 is a very useful tool in a biological laboratory. It is used, for instance, routinely in medical laboratories in "centrifuging" blood for the purpose of separating the blood cells from the liquid portion.

One has succeeded in developing so-called *ultracentrifuges* which rotate at such a rapid rate that a centrifuged particle may acquire a d'Alembert weight about a million times its normal terrestrial weight. At such high frequencies of revolution sugar can be precipitated from its solutions. A limit is set to the achievable rotational speed by the fact that the material of which the centrifuge is made is torn to pieces by the action of excessively high inertial forces.

Svedberg used the ultracentrifuge for the purpose of separating different proteins, the essential building stones of living organisms, from each other and was able to determine their molecular weight by measuring their rate of sedimentation in the ultracentrifuge. Some of the proteins turned out to have molecular weights of over a million. Svedberg's method is of great importance, since the separation and identification of proteins by chemical methods are often difficult or impossible.

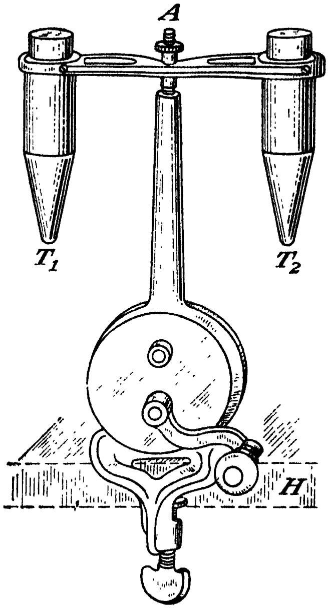


FIG. 6.5. A hand-driven centrifuge. The tubes T_1 and T_2 are rotated about axis A by turning the crank H. The tubes contain the liquid to be centrifuged. They assume a nearly horizontal position when revolving. (*Courtesy of Central Scientific Company.*)

A very interesting application of the centrifuge to the study of living cells was made by Harvey, who designed a centrifuge microscope which makes it possible to observe living cells as they are being centrifuged and thus subjected to conditions equivalent to a many thousandfold increase of gravitational forces. Figure 6.6 taken from a paper by Harvey shows how the different materials contained in the cell are separated into layers according to their density. The lightest material

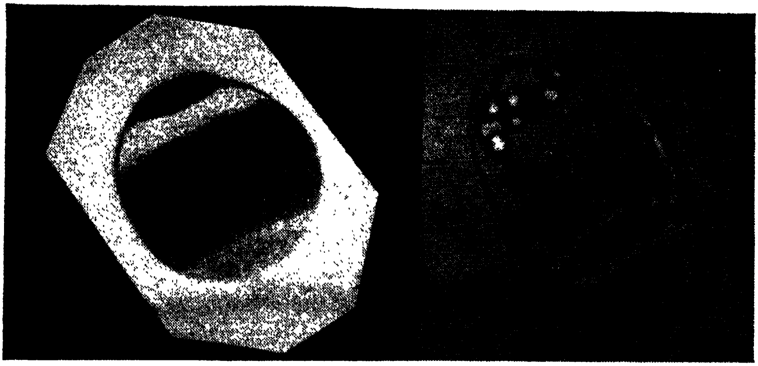


FIG. 6.6. *Left:* Centrifuged sea-urchin (Tripneustes) egg showing oil, clear yolk, and mitochondria layers. *Right:* A worm (Nereis) egg centrifuged in an air turbine at 84,000 × gravity, showing oil globules, a large nucleus with small nucleolus inside, a band of large yolk granules and very fine granules below. (*From Harvey, J. Applied Phys. Vol.* 9, 1938.)

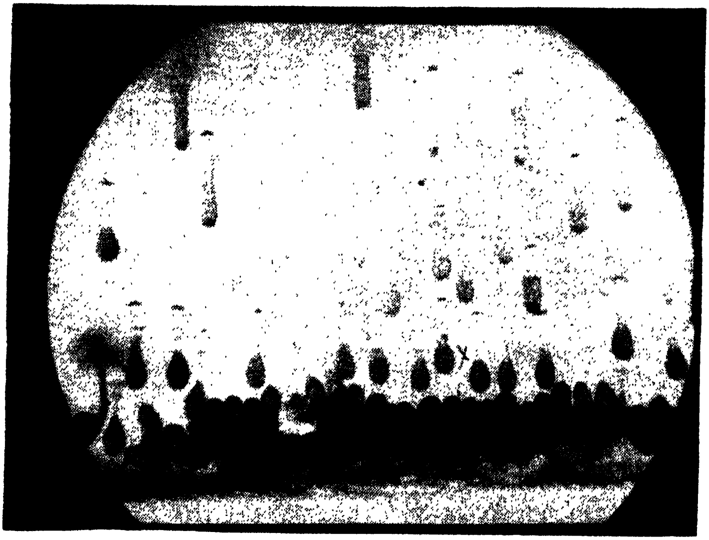


FIG. 6.7. A frame of a motion-picture film taken through the centrifuge microscope showing sea-urchin eggs being pulled into two halves by inertial forces (5,500 × gravity). 0.8-sec exposure; revolving at 7,500 rpm. (*From Harvey, J. Applied Phys. Vol.* 9, 1938.)

(oil) is seen to form a pole near the northwest corner of the left photograph, whereas the heaviest material is sedimented at the opposite end of the cell. The forces acting on the light pole and the heavy pole of the cell are directed toward the axis of rotation and away from it, respectively, and thus tend to elongate the cell. Figure 6.7 shows how sea-urchin eggs are deformed and, as a result of the elongation, pinched off into a light and heavy "half" (see cells to the right of the cross). These different halves can, for instance, be fertilized, and their development can be studied. At still higher speeds of revolution the "halves" break in two, forming "quarters." The varied, interesting studies on the physical and biological properties of cells made possible by this method will be found discussed in Harvey's paper (see bibliography).

QUESTIONS AND PROBLEMS

1. If you were to observe, while in a sealed box, all objects to fall with constant acceleration in a given direction, how could you decide whether or not this behavior was due to gravitation or merely due to the acceleration of the box?

2. Two physicists are traveling to the moon in a rocket which is equipped with laboratory facilities. After all the fuel has been used up, in other words while the rocket continues moving owing to its acquired momentum, they decide to repeat Galileo's experiments on the motion of bodies down an inclined plane. How do their results compare with those of Galileo?

3. An enterprising elevator operator of a busy skyscraper opens a small grocery store in his elevator. He weighs his groceries by means of a spring scale, and he sells them while his elevator is accelerating upward and charges according to the indicated weight. When the goods are delivered to him, he checks their weight while the elevator is retarding on its way up and pays according to the recorded weight.

a. How much profit will he make on a pound of cheese if he buys it at 20 cents per pound and sells it at the same price? (Assume the acceleration in all cases to be 8 ft/sec^2.)

b. What kind of scale will the police department suggest for this enterprise?

4. How fast must a car accelerate in order for a passenger to be pressed against the back wall with a force equal to twice his weight?

5. What would happen to the gas bubbles if you were to centrifuge ginger ale?

6. If your hand were bitten by a snake, what could you do as an alternative to sucking out your wound?

7. If the box of Fig. 6.3 is accelerated at the rate of 200 cm/sec^2, what d'Alembert weight will the spring scale S indicate for a body W having a mass of 2 kg?

8. Determine the per cent increase (or decrease) in the scale reading for a 6-kg body which rests on a spring scale in an elevator under the following conditions:

a. The elevator moves at the constant speed of 2 m/sec (1) upward, (2) downward.

b. The elevator moving initially upward with the speed of 6 m/sec suffers a change in speed (1) to twice the original value, (2) to half the original value in 2 sec.

c. Answer the questions in part *b* assuming the elevator to be moving originally downward.

9. Could the division of cells into halves and quarters observed by Harvey in a centrifuge occur also on the surface of a sufficiently heavy and dense planet? Explain in detail the reason for the "pinching off" of cells observed by Harvey. In what fashion are the ingredients of a cell "sorted" in the process of centrifugation?

BIBLIOGRAPHY

On ultracentrifuge:

Beams, J. W., *Science in Progress*, ed. by G. A. Baitsell, p. 232, 1940.

On centrifugation of eggs (containing an extensive bibliography):

Harvey, E. N., *J. Applied Phys.*, Vol. 9, pp. 68–80, 1938.

CHAPTER 7

DYNAMICS: 4. VECTORS AND THE GENERALIZATION OF THE CONCEPT OF ACCELERATION

a. What dynamic manifestations of centripetal acceleration can you name?

b. Is the centripetal acceleration still present if the body is not revolving uniformly?

c. Is it reasonable to assume that an object which does not change its distance from the center of revolution is accelerated toward the center?

d. In what way does the procedure of adding vectors differ from the addition of scalars?

e. What does it mean to subtract vector $\vec{A}$ from vector $\vec{B}$?

f. Can a point have five different velocities in different directions at the same time? If the answer is yes, how can you find its resultant motion?

g. Does the validity of vectorial rules of addition for acceleration vectors imply their validity for force vectors?

h. Do the same rules apply to momentum vectors? Why? Or why not?

i. On what factors does the centripetal acceleration of a revolving body depend? The centripetal force? The centrifugal force?

j. What happens to the centripetal acceleration of a revolving body if you double the orbital velocity v and halve the angular velocity ω?

k. If you know the mass of a planet describing a circle of known radius with a given period, can you find the force that is acting on it? How?

l. If the force of attraction in *k* were to cease suddenly, how would the planet continue to move?

m. Is there some simple relationship between the orbital velocity and centripetal acceleration in circular motion?

n. How could you visualize the falling motion of the moon toward the earth?

In Chap. 6 we have been led by dynamical considerations to the assumption of a centripetal acceleration in the case of a body moving in a circular path at a constant rate. This assumption seems contrary to our common sense, since the body, which is supposedly accelerated toward the center, evidently does not move toward it! In order to resolve the apparent contradiction, we must understand the general meaning of the concept of acceleration. This, however, requires familiarity with the properties of vector quantities. We shall, therefore, interpose at this point a brief discussion of vector addition and subtraction.

We shall derive the rules of vector addition from the consideration of the simplest kind of vector we can imagine, the displacement vector, since the procedure can be easily visualized and appears to be intuitively self-evident.

Vector Notation

If a kangaroo jumps from A to B, we shall symbolize its displacement by an arrow drawn from A to B: $A \rightarrow B$. The displacement BA will be symbolized by the opposite arrow: $A \leftarrow B$. We shall define $\overrightarrow{AB} = -\overrightarrow{BA}$; that is, *a negative sign in front of a vector symbolizes reversal of its direction.* We can draw in the proper direction a small arrow which represents the displacement of the animal according to some appropriately reduced scale. We can then operate with it, using it as a symbol of the actual displacement of the kangaroo.

Addition of Displacement Vectors

It is easy to see that vectors do not, in general, add up in a simple algebraic manner. If the kangaroo jumps 4 ft east and then 3 ft north, its resultant displacement R from the origin is not 7 ft but, as we can easily see from the diagram of the actual motion (Fig. 7.1), $R = \sqrt{4^2 + 3^2} = 5$ ft. We can obtain the **resultant displacement** vector by first drawing the two **component vectors** so that the tail of the second vector is placed against the head of the first vector and then joining the tail of the first to the head of the second vector. The arrow which we thus obtain symbolizes the resultant displacement of the jumper from his origin A to his final destination C.

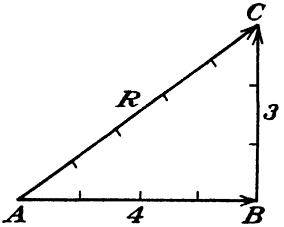


FIG. 7.1. Addition to two vectors: $\overrightarrow{AB} + \overrightarrow{BC} = \overrightarrow{R}$.

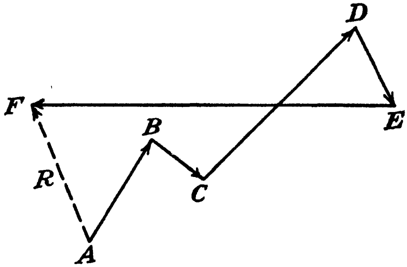


FIG. 7.2. Addition of five vectors: $\overrightarrow{AB} + \overrightarrow{BC} + \overrightarrow{CD} + \overrightarrow{DE} + \overrightarrow{EF} = \overrightarrow{R}$.

The same procedure can be extended to a complicated sequence of displacements (Fig. 7.2): As before, we draw a *chain of vectors* so that the *tail of any vector is placed against the head of the preceding vector;* **the resultant vector** *is then obtained by joining the tail of the first to the head of the last vector.*

Subtraction of Displacement Vectors

The rules for subtraction of displacement vectors follow directly from the above rules for addition. Since $\overrightarrow{AB} = -\overrightarrow{BA}$, we can transform the problem of subtraction into one of addition as follows: $\vec{a} - \vec{b} = \vec{a} + (-\vec{b})$, that is, the subtraction of a vector b is equivalent to the addition of the vector $-b$. The graphical determination of the vectors $\vec{s} = \vec{a} + \vec{b}$ and $\vec{d} = \vec{a} - \vec{b}$ is shown in Figs. 7.3*a* and *b*.

We can also show that $\vec{b}$ is the vector difference between $\vec{s}$ and $\vec{a}$: We see from Fig. 7.3*a* that $\vec{a} + \vec{b} = \vec{s}$. Hence, $\cancel{\vec{a}} + \vec{b} - \cancel{\vec{a}} = \vec{s} - \vec{a}$ or $\vec{b} = \vec{s} - \vec{a}$. This suggests another construction of the **difference between two vectors** (in this case between $\vec{s}$ and $\vec{a}$). *Place the two vectors tail to tail and join their tips.* The vector joining their tips as indicated in Fig. 7.3*a* is their vector difference.

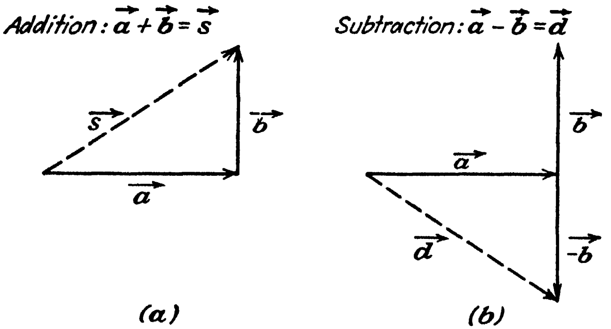

Addition: $\vec{a} + \vec{b} = \vec{s}$
Subtraction: $\vec{a} - \vec{b} = \vec{d}$
$\vec{s}$
$\vec{b}$
$\vec{a}$
$\vec{a}$
$\vec{b}$
$\vec{d}$
$-\vec{b}$
(a)
(b)

FIG. 7.3*a*. Addition of two vectors: $\vec{a} + \vec{b} = \vec{s}$. *b*. Subtraction of a vector: $\vec{a} - \vec{b} = \vec{d}$.

Extension to Velocity Vectors

The above procedure as applied to vectors representing consecutive displacements appears self-evident. Does it also apply to simultaneous displacements? This amounts to the question: If a passenger on a train steps 3 ft toward the north wall of the train while the train rolls 12 ft eastward, is the resultant displacement the same as in the case when the train moves first and stops whereupon the passenger takes a step? Is the result different if the motion of the passenger precedes that of the train? The answer is that the sequence of displacements is immaterial and the result of consecutive displacements is identical with that of simultaneous motion. Hence, the rules which apply to the addition of consecutive displacements will also hold for simultaneous ones.

Now let us pause to consider what the preceding statement implies. Suppose that in our previous example the time required for the train to roll 12 ft while the passenger moves 3 ft is 1 sec. Then the respective velocities are 12 ft/sec east and 3 ft/sec north. In this case the velocity vectors are represented by arrows of the same magnitude and direction as the displacement vectors, *velocity vectors being simply vectors of displacement per unit time.* Hence, *the rules which apply to the addition of displacement vectors also apply to velocity vectors.* (If the velocity is not constant, we can choose a conveniently small unit of time during which the velocity change is negligible. The displacement during that short time interval will be proportional to the instantaneous velocity and will have the same direction.)

Extension to Acceleration

The extension of our rules of vectorial addition to acceleration vectors follows from their validity for velocity vectors in the same way that the applicability of the rules to velocity vectors was derived from a consideration of simultaneous displacements.

Our original definition equation for the average acceleration in linear motion was

$$a = \frac{v_2 - v_1}{t}$$

If we don't restrict ourselves to the consideration of motion in a constant direction, we shall admit that the direction of $\vec{v}_2$ may not be the same as that of $\vec{v}_1$. We shall then generalize our definition equation of acceleration by using vector notation:

$$\boxed{\vec{a} = \frac{\vec{v}_2 - \vec{v}_1}{t}} \tag{7.1}$$

Figure 7.4 shows the velocity vectors v_1 and v_2 at the points P_1 and P_2 for a body moving on a curved path.

The only vector quantities on the right side of equation (7.1) are $\vec{v}_1$ and $\vec{v}_2$ (t being a scalar). Thus, we see that the direction of $\vec{a}$ is given by the direction of the vector difference $\vec{v}_2 - \vec{v}_1$, which is represented graphically in Fig. 7.4*b*.

By setting t equal to one unit of time in equation (7.1), we see that the acceleration vector $\vec{a}$ appears as the vectorial velocity difference per unit time. Rules of addition which are valid for velocities obviously apply as well to velocity differences (since any velocity can be represented as a difference between two velocity vectors). Hence, acceleration vectors may be added and subtracted just as velocity vectors according to the rules established above.

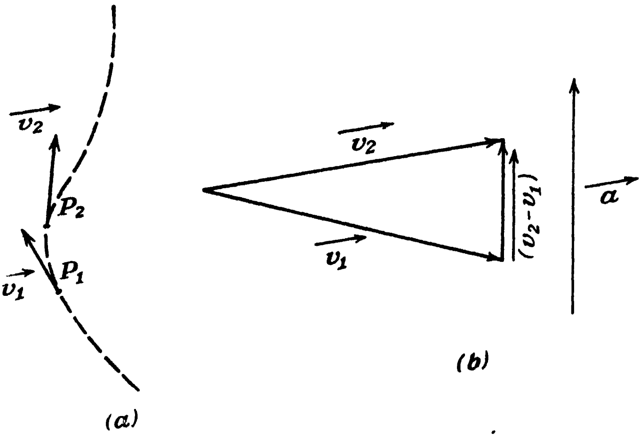


FIG. 7.4*a*. Variation of velocity in curvilinear motion. *b*. Vector diagram of velocity change: $\vec{a}$ is parallel to $(\vec{v}_2 - \vec{v}_1)$.

Extension to Forces

Since $\vec{F} = m \cdot \vec{a}$ (where m is a scalar), the direction of the force vector will always coincide with the direction of the corresponding acceleration vector. The same rules which hold for acceleration vectors apply then to the force vectors which are proportional to them.

Extension to the General Case of Any Kind of Vectors

By similar reasoning, the extension of the rules of vector manipulation may be carried out for other vector quantities in physics, examples of which will be encountered later on. In general, we define as a vector any quantity which obeys the vectorial rules of computation which are valid for forces, velocities, etc.

Acceleration in Circular Motion

In circular motion it seems difficult to comprehend that a body may have an acceleration $\vec{a}$ different from zero although the magnitude of its velocity may not change. In symbolic notation we can express it as follows: $|\vec{v}_2| = |\vec{v}_1|$ but $|(\vec{v}_2 - \vec{v}_1)| \neq 0$. Since $\vec{v}_2 - \vec{v}_1$ is not zero, it follows from (7.1) that $\vec{a}$ is also different from zero! *Hence, the motion is an accelerated one by definition.* Figure 7.5*a* shows the different velocities of a point passing at constant speed points P_1 and P_2. Figure 7.5*b* shows the velocity change $(v_2 - v_1)$ which, as we shall see presently, is the consequence of the centripetal acceleration of the moving point.

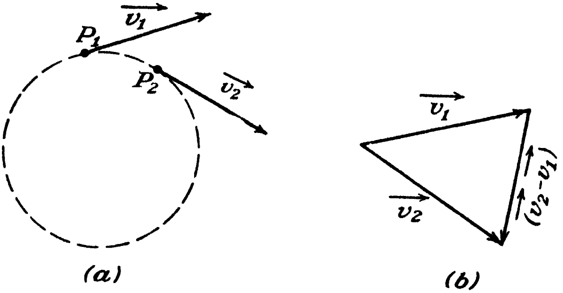


FIG. 7.5*a*. Variation of velocity in uniform circular motion. *b*. Vector diagram of velocity change.

The General Case of Curvilinear Motion

The validity of results obtained by an analysis of circular motion is not limited to bodies moving on the periphery of a circle. The results may be applied to objects moving along other types of curves as well, since at any instant any short portion of the curve traversed by the body may be approximated by a section of a circle which best fits the curve at that point and whose radius (radius of curvature) can be used for purposes of computation according to formulas derived below.

Magnitude of Centripetal Acceleration

On the basis of Newton's second law $(F = ma)$, we can evaluate the force acting on a given mass moving in a straight line if we determine its acceleration with a yardstick and a stop watch. Is a similar determination of F possible in the case of circular motion? What do we have to measure in order to find a? As we shall see presently, we have merely to measure the radius of curvature R and the magnitude of the peripheral velocity to obtain the centripetal acceleration $a = v^2/R$.

The expression for the *magnitude* of $\overrightarrow{a}$ can be derived as follows: Consider a mass m moving uniformly along the periphery of a circle. Although the velocity vector changes its direction, its magnitude remains constant. This is designated by the notation: $|\overrightarrow{v_1}| = |\overrightarrow{v_2}| = v =$ constant, whereas $\overrightarrow{v_1} \neq \overrightarrow{v_2}$. The distance s covered in t seconds by m along the circular arc between the points 1 and 2 (Fig. 7.6*b*) is then given by

$$\boxed{s = vt} \tag{7.2}$$

The radius R connecting m to the center of the circle changes its direction by the angle θ. It is evident from Fig. 7.6*b* that the direction of v is also changed by the same angle θ.

The angular measure of θ [see equation (3.3)] is

$$\theta = \frac{s}{R} \text{ radians} \tag{7.3}$$

Substituting (7.2) in (7.3), we get

$$\theta = \frac{s}{R} = \underline{\frac{vt}{R} = \theta} \tag{7.4}$$

Now, Fig. 7.6*a*, where the two vectors $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ have been transferred parallel to themselves from Fig. 7.6*b*, shows that the direction of v also changed by θ radians. The line joining the end points of $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ represents the vector difference $\overrightarrow{\Delta v} = \overrightarrow{v_2} - \overrightarrow{v_1}$. This line is obviously shorter than the circular arc (dotted line) described by the radius $R = v$ about apex O in Fig. 7.6*a*. But the difference in length between this line and the arc will diminish more and more as we make θ progressively smaller. To make θ small amounts physically to making our two successive observations of v in very quick succession, *i.e.*, to choosing a very small time interval t. In that case we can treat Δv as if it were equal to the corresponding circular arc and express θ from Fig. 7.6*a* with a high degree of approximation as follows:

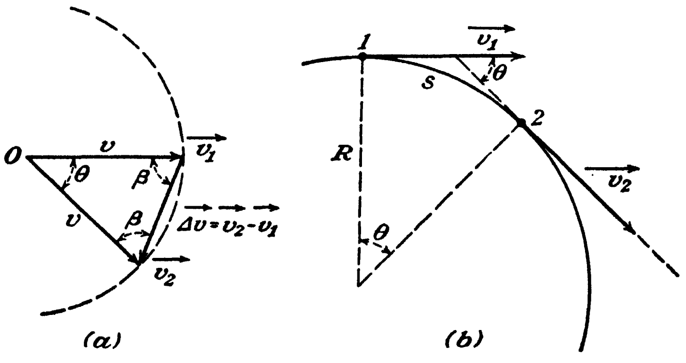


FIG. 7.6.

$$\theta = \frac{\Delta v}{v} \tag{7.5}$$

Now, $\overrightarrow{\Delta v} = \overrightarrow{v_2} - \overrightarrow{v_1}$ and from (7.1):

$$\vec{a} = \frac{\vec{v}_2 - \vec{v}_1}{t} = \frac{\overrightarrow{\Delta v}}{t}$$

hence,

$$\Delta v = at \tag{7.1a}$$

Substituting (7.1*a*) in (7.5), we get

$$\underline{\theta = \frac{\Delta v}{v} = \frac{at}{v} = \theta} \tag{7.6}$$

We now have two different expressions for θ, (7.4) and (7.6), which we can equate to each other:

$$\theta = \frac{at}{v} = \frac{vt}{R}$$

We obtain thus

$$\boxed{a = \frac{v^2}{R}} \quad \text{(the \textbf{centripetal acceleration})} \tag{7.7}$$

This result can be also rewritten in terms of the angular velocity ω and R, using the relation $v = \omega R$ [equation (3.4)]. Substituting this in (7.7), we get

$$\underline{a = \omega^2 R} \tag{7.7a}$$

or we can also write an expression symmetrical in v and ω from (7.7*a*):

$$\underline{a} = \omega^2 R = \omega(\omega R) = \underline{\omega v} \tag{7.7b}$$

(It is sufficient to remember only equation (7.7) and the relation $v = \omega R$.)

We have derived analytically the magnitude of the centripetal acceleration, but we have not shown by analysis what the direction of this vector should be. Our contention that it should point toward the center of revolution was, so far, based on inference from dynamical observations (centrifugal force interpreted as a d'Alembert force resulting from centripetal acceleration). As is shown in Fig. 7.6*b*, the instantaneous velocity vector is tangential to the circular orbit of motion. From Fig. 7.6*a* we can readily see that the base angles β of the isosceles triangle approach 90° as θ diminishes toward zero (which happens as we choose progressively smaller values of the observation time t). For the limiting case of an infinitely small t, θ will be indistinguishably close to 0 and β to 90°, so that we can say that for a very short time t of observation *the vectorial velocity increment* $\overrightarrow{\Delta v}$ *will be perpendicular to the* $\vec{v}$ *vector.* Since $\vec{v}$ is tangential to the circle (see Fig. 7.6*b*), $\overrightarrow{\Delta v}$ will be radial pointing toward the center of the circle.

Since the directions of $\vec{a}$ and $\overrightarrow{\Delta v}$ are the same, as can be inferred from equation (7.1) ($\vec{a} = \Delta\vec{v}/t$), we can state that *the acceleration of a point moving uniformly in a circular orbit is a centripetal acceleration, being perpendicular to the tangent of the circle, or, which is the same, to the circular orbit itself and hence pointing along the radius toward the center of curvature.*

Visualization of the "Falling" Process Connected with Centripetal Acceleration

Our final task is now to resolve the apparent contradiction to common sense which we cited at the outset of this discussion. Suppose the moon moves uniformly around the earth in a circle. (This is only an approximation.) This implies that the moon is accelerated constantly toward the earth with a centripetal acceleration, $a = v^2/R$. According to Newton's second law, this centripetal acceleration is due to a centripetal force evidently exerted upon the moon by the earth which is at the center of its orbit. But we see no obvious evidence of this acceleration. The moon does not appear to fall toward the earth! Strange as it may seem, however, the moon does fall toward the earth (while at the same time the earth falls toward the moon). This can be visualized as follows (Fig. 7.7): Imagine the moon at point 1 of its orbit.† Consider its motion for the next t hours. If no gravitational forces were exerted by the earth on the moon, the latter would continue moving straight toward point 2 in accordance with Newton's first law. This would remove it farther and farther away from the earth (distance $E - 2 > E - 1$). Since actually t hours later the moon is known to appear at 3 rather than 2 (keeping about constant distance from the earth), some process must have brought it closer to us than it would have been in the absence of interaction with other bodies. We are justified in calling this process of approximation toward the center of the earth "falling." While it is difficult to visualize simultaneously how

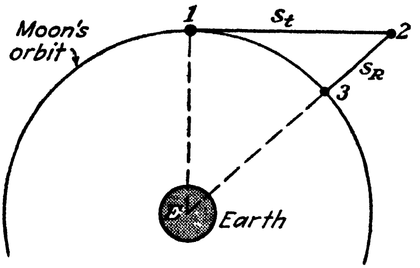


FIG. 7.7. Centripetal acceleration of the moon.

---

† The moon is not revolving about the center of the earth as is shown in Fig. 7.7 for simplicity. Both bodies are actually revolving about a common "mass center," which, however, is quite close to the center of the more massive earth. The same holds for the revolution of planets about the sun and of double stars about each other. Thus, both interacting bodies revolve, and each experiences a different centripetal acceleration which, when multiplied by the mass of the body, gives the same centripetal force for both bodies (Newton's third law).

the moon tends to move in a straight line tangentially while also falling toward the earth radially, it can be pictured quite simply if we separate these two processes in time: First imagine the moon to move for t hours with terrestrial gravitation "switched off," then imagine the moon (which has arrived at 2 by this time) to follow the gravitational pull of the earth, no longer obeying the law of inertia (except in radial motion). This will get it to point 3. The interesting point about this circular motion is that the centripetal acceleration is of such a magnitude that the moving body will fall radially in a given time through the same distance S_R as it would have receded from E (moving along the line 1-2) in the absence of a centripetal attraction. Thus, the distance from the center of revolution remains constant and the existence of an acceleration is not apparent to a naïve observer.

Centripetal Force

We ascribe the centripetal acceleration of a body to the action of a centripetal force. Once the centripetal acceleration is known [equation (7.7)], the **centripetal force** can easily be determined on the basis of Newton's second law:

$$F = ma = \boxed{m\frac{v^2}{R} = F} \tag{7.8}$$

F is given in dynes when m is measured in grams, v in centimeters per second, and R in centimeters. It is given in pounds when m is measured in slugs, R in feet, and v in feet per second.

Centrifugal Force

"To every action there is an equal and opposite reaction": A centripetal force F_1 compels the stone which we whirl about in a circle at the end of a string to describe its circular path (Fig. 7.8). This force is exerted upon stone S through the string by the center pin C. According to Newton's third law, an equal and opposite reaction is exerted by the stone upon the pin C. This force seemingly tends to pull C outward from its central position. It is called the **centrifugal force.** It is not to be confused with the centrifugal inertial force discussed in Chap. 6 experienced by a revolving observer. It is to be kept in mind that the centrifugal force referred to here is the *reaction* exerted by the revolving body S upon the center of revolution C, which exerts the centripetal force upon S, thus compelling it to move in a circle.

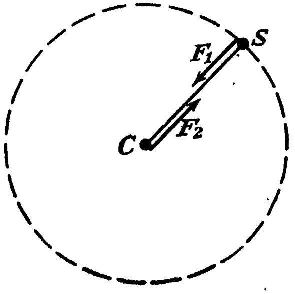


FIG. 7.8. Centripetal and centrifugal force.

Centrifugal D'Alembert Force. There is frequent confusion in the literature on the use of the term *centrifugal force.* It is often used to designate what we called the centrifugal d'Alembert force. The latter is a fictitious quasi-gravitational force introduced by an observer in a rotating system who wishes to interpret static observations in his system considering his system as being at rest. For instance, in discussing sedimentation of blood cells in a centrifuge, we can consider the centrifuge as not rotating if we ascribe to each particle a "centrifugal d'Alembert weight" equal to its mass times centripetal acceleration. The preferred terminology is to designate as "centrifugal force" the reaction to centripetal force.

APPENDIX

Equilibrium of a Point

A point is said to be in equilibrium when the vector sum of forces acting upon it is zero.

How can we determine whether or not a point subject to the given forces (Fig. 7.9*a*) is in equilibrium?

Method I. Graphical Method

Form a vector chain (Fig. 7.9*b*). The point is in equilibrium when the resultant R is zero. In this case the tip of F_4 would touch the tail of F_1. Thus we see that the criterion for the equilibrium is the closing of the vector chain. **A point is in equilibrium if the forces acting upon it add vectorially to form a closed polygon.**

Method II. Resolution Method

This method is analytical, depending on computation rather than on an accurate drawing, as method I. Since any force can be conceived of as being the resultant of two forces, we can imagine the oblique forces F_2 and F_4 as being composed of two *components* each. We choose these components parallel to two conveniently assumed axes. In most cases it is best to choose these axes perpendicular to each other, such as the X and Y axes in Fig. 7.9*c*. Thus we obtain for F_2 the rectangular components $(F_2)_x$ and $(F_2)_y$ and, for (F_4), $(F_4)_x$ and $(F_4)_y$.

Since the effect of the sum $\overrightarrow{(F_2)_x} + \overrightarrow{(F_2)_y}$ is equivalent to the effect of $\overrightarrow{(F_2)}$, we replace $\overrightarrow{F_2}$ by its two components $(F_2)_x$ and $(F_2)_y$ in Fig. 7.9*c* and similarly the force F_4 by its X and Y components. We thus obtain a system of vectors which are parallel either to the X axis or to the Y axis. All the "X vectors" can be added algebraically: $F_3 + (F_4)_x - (F_2)_x = R_x$.† R_x is the X component of the resultant vector. Similarly, we obtain for the "Y vectors": $-F_1 - (F_2)_y + (F_4)_y = R_y$.

From this the magnitude of the resultant can be computed (Fig. 7.9*d*):

$$R = \sqrt{R_x^2 + R_y^2}$$

---

† We have taken vectors pointing toward the right along the X axis and upward along the Y axis as positive, assigning the negative sign to opposite vectors. This specifying of signs is arbitrary and can be equally well reversed.

The direction of the resultant can be determined by the angle against the vertical (Fig. 7.9*d*): $\tan\theta = R_x/R_y$. The magnitudes of $(F_2)_x$ and $(F_2)_y$ as well as the components of F_4 (which are needed for the above computations) are found from Fig. 7.9c as follows: $\sin\alpha = (F_2)_y/F_2$.

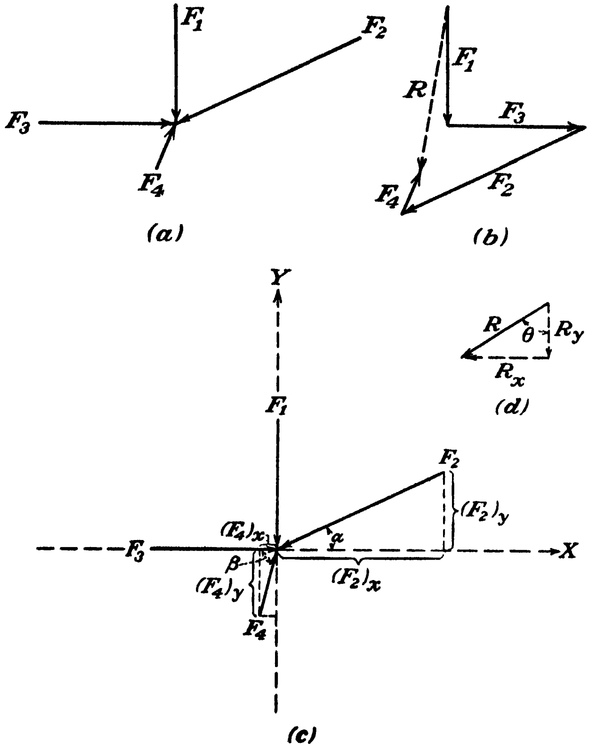


FIG. 7.9*a*. Forces acting on a point. *b*. The resultant of the forces of Fig. 7.9*a*. *c*. Resolution into rectangular components of the forces of Fig. 7.9*a*. *d*. Determination of the resultant according to the resolution method.

Hence, $$(F_2)_y = F_2 \sin\alpha$$

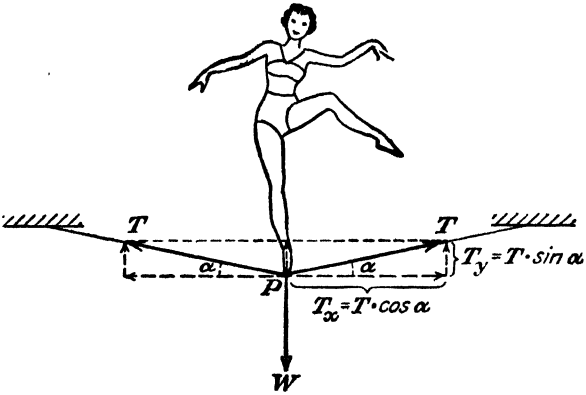
Also, $$\cos\alpha = \frac{(F_2)_x}{F_2}$$

hence, $$(F_2)_x = F_2 \cos\alpha$$

Similarly we find $(F_4)_y = F_4 \sin\beta$ and $(F_4)_x = F_4 \cos\beta$. The point is in equilibrium when $R = 0$. Equilibrium exists if we can show that the components of R are zero. **Thus, a sufficient condition for the equilibrium of the point is $R_x = 0$ and $R_y = 0$.**

Example 1

How large is the tension in the rope under the tightrope walker in the following figure? Assume that her weight is 120 lb and $\alpha = 15°$.



Solution. We know that the point P of the rope is in equilibrium. The following forces act upon that point: W, the weight of the rope walker, and the two tension forces T transmitted through the string to the point P. The sum of the X components of the tensions T is zero, since both are equal to $T \cos \alpha$ pointing in opposite directions:

$$\Sigma T_x = T \cos \alpha - T \cos \alpha = 0$$

Since P is in equilibrium, the sum of the Y components must also be zero·

$$2T_y - W = 2T \sin \alpha - W = 0$$

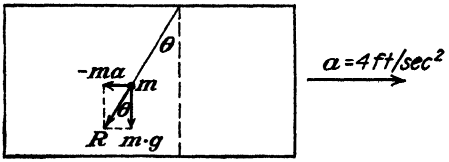
hence,

$$T = \frac{W}{2 \sin \alpha} = \frac{120}{2(0.26)} = 230 \text{ lb}$$

We see that the tension in the rope is larger than the weight of the rope walker in our example. It is seen from $T = W/(2 \sin \alpha)$ that, as α approaches 0° and $\sin \alpha$ the value 0, T goes toward ∞. Hence, to avoid excessive tension, which may break the rope, it should not be too tight!

Example 2

What will be the direction of the plumb line in a train which is accelerating at the rate of 4 ft/sec^2?

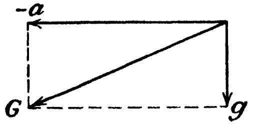


Solution. The problem is solved easily with the aid of d'Alembert's principle. All masses inside the accelerated train possess, in addition to their weight (mg), the horizontal "d'Alembert weight" ($-ma$). The resultant force acting on any body is the vector sum $\overrightarrow{R} = \overrightarrow{mg} + (-\overrightarrow{ma})$. This vector addition is carried out in our diagram. We see that the resultant force is not vertical. Bodies which are suspended in the train hang parallel to the direction of $\overrightarrow{R}$. Bodies which are released fall also in the same direction. This is easy to see if, instead of adding vectorially $(\overrightarrow{-ma})$ and $\overrightarrow{mg}$, we add the vectors $\overrightarrow{-a}$ and $\overrightarrow{g}$. We obtain a resultant acceleration vector $\overrightarrow{G}$ which is parallel to $\overrightarrow{R}$. The direction of R is easily obtained as follows: We see from the figure that $\tan\theta = ma/mg = a/g = \frac{4}{32} = \frac{1}{8}$; hence, $\underline{\theta = 7°}$.

If we wish, we can also compute the resultant apparent weight of the body:

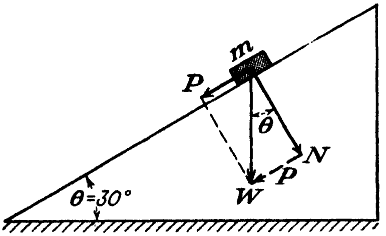
$$R = \sqrt{(mg)^2 + (ma)^2} = m\sqrt{g^2 + a^2} = m\sqrt{32^2 + 4^2} = 32.2\ m$$

The expression $\sqrt{g^2 + a^2} = 32.2$ is the resultant acceleration G. We see that it is somewhat larger than g, in addition to having a different direction. The weight of bodies registered by a spring scale is correspondingly increased.



Example 3

A mass m slides down an inclined plane shown in the figure. How large is (*a*) the accelerating force, (*b*) the "normal force" N which the mass exerts at right angles to the inclined plane, (*c*) the acceleration of the mass?



Solution. *a.* We resolve the weight $\overrightarrow{W} = \overrightarrow{mg}$ into two components: $\overrightarrow{P}$ parallel to the inclined plane and $\overrightarrow{N}$ perpendicular to it. We recognize from similarity of triangles that the angle between $\overrightarrow{W}$ and $\overrightarrow{N}$ is θ. From the diagram,

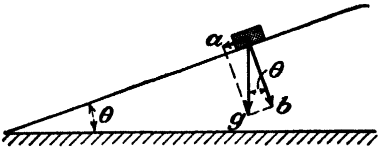
$$\sin\theta = \frac{P}{W} \qquad \text{or} \qquad P = W\sin\theta = W\sin 30° = \underline{0.5W}$$

b. Similarly, $\cos\theta = N/W$ or $N = W\cos 30° = \underline{0.867W}$.

c. Since the body is constrained to move parallel to the inclined plane, only the force component $\overrightarrow{P}$ will contribute to its acceleration. In this case Newton's second law will assume the form

$$P = W\sin\theta = ma \qquad \text{or} \qquad a = \frac{W\sin\theta}{m} = \frac{mg\sin\theta}{m} = \underline{g\sin\theta}$$

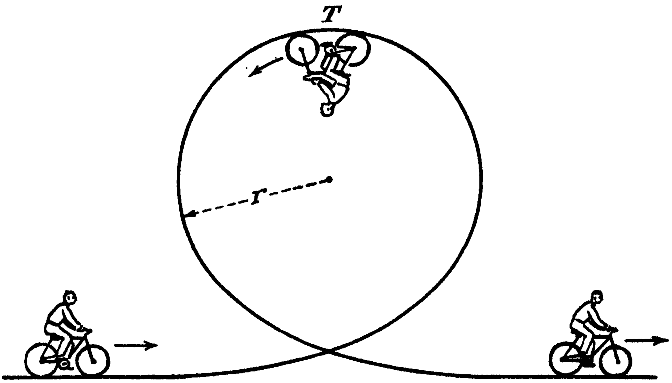
The same value for the acceleration down the inclined plane is obtained if we resolve the acceleration of gravity into two components, one parallel to the inclined plane (a) and the other perpendicular (b) as shown in this figure:



We see from the diagram $\sin \theta = a/g$, $a = g \sin \theta$.

Example 4

A circus performer intends to "loop the loop" on a bicycle as illustrated in the figure. If he maintains a constant speed, what is the minimum speed he must have in order not to lose contact with the loop when he reaches the top of the loop? (Assume a loop of radius $r = 8$ ft.)



Solution. In the absence of gravity we should have to consider only one force, the centripetal force necessary to keep the cyclist on the circular orbit by imparting to him the centripetal acceleration. This centripetal force is exerted by the loop upon the tires of the bicycle and points toward the center of the circular loop.

Now, in actual practice when the cyclist is in position T, the force of gravity provides part of the centripetal force, the other part being provided by the reaction R of the loop. This is expressed by the following equation:

$$m\frac{v^2}{r} = R + W \tag{1}$$

(where v^2/r is the centripetal acceleration; R, the force of reaction exerted by the loop upon the bicycle; and $W = mg$, the weight of the cyclist plus the bicycle.) From this equation follows:

$$R = m\frac{v^2}{r} - mg = m\left(\frac{v^2}{r} - g\right) \tag{2}$$

We see that the reaction of the loop is equal to the difference between the centripetal force and the weight of the moving body. As long as the centripetal acceleration

$v^2/r > g$, there is a reaction and hence a contact between the loop and the cycle. For the value of v at which $v^2/r = g$, the reaction R becomes zero, and hence we might as well remove the loop at this moment, since it serves no purpose. At this point the entire centripetal force is provided by gravity. We can thus say that the cyclist loses contact with the loop in his top position when $v^2/r = g$. Hence, the minimum speed required is $v = \sqrt{rg} = \sqrt{8(32)} = \underline{16\text{ ft/sec.}}$

QUESTIONS AND PROBLEMS

1. A system of moving bodies has the following masses and velocities: $m_1 = 20$ gm, $v_1 = 100$ cm/sec 30° north of east; $m_2 = 10$ gm, $v_2 = 50$ cm/sec 60° south of west; $m_3 = 50$ gm, $v_3 = 350$ cm/sec north; $m_4 = 500$ gm, $v_4 = 100$ cm/sec east. Find the resultant momentum of this system.

2. An airplane is flying south at the rate of 300 mi/hr. The wind is blowing westward with the velocity of 60 mi/hr. Find analytically and graphically the velocity of the airplane relative to the ground.

3. A train approaches a railroad station moving westward with the speed of 6 mi/hr. A waiter in the dining car happens to be walking northward with the speed of 3 mi/hr while a fly located on his bald head is creeping 30° north of east with the speed of 0.2 mi/hr. Find graphically the velocity of the fly relative to the ground.

4. An observer in a train which is accelerated northward at the rate of 8 ft/sec^2 releases a ball. In what direction does the ball move? How large a distance does it traverse during the first half second?

5. Food is propelled in our intestines by peristaltic motion. Show by a vector diagram how peristalsis produces a force which propels food along the intestine.

6. Deduce the law of elastic reflection for a ball striking a wall at oblique incidence. (Hint: Resolve the momentum of the ball into two suitably chosen components.)

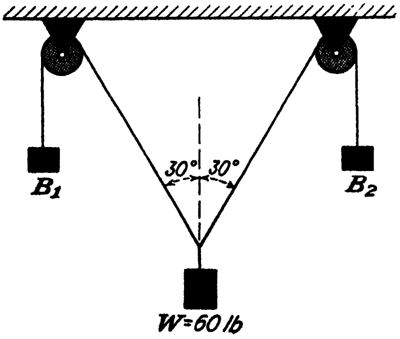
7. A 10-lb body rests on a frictionless plane which is inclined 30° against the horizontal level.

a. How large a force must be applied parallel to the inclined plane in order to keep the body in equilibrium?

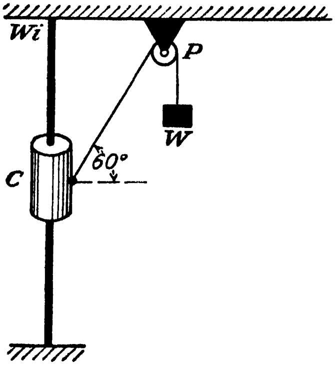
b. How large is the component of the body weight at right angles to the inclined plane?

c. How far would the body slide along the inclined plane in 2 sec if allowed to move freely?

8. A 60-lb body is suspended from pulleys as shown in the figure. How large are the two equal weights B_1 and B_2?



9. The cylinder *C* (10 lb) slides freely up and down the vertical wire *Wi*. How large must be the weight *W* in order to maintain *C* in equilibrium in the setup shown in the diagram?



10. A flowerpot is placed on the rim of a merry-go-round of radius $R = 50$ cm which is rotating about a vertical axis at the rate of 6 revolutions per second. In what direction (give angle against the vertical) does the stem of the developing seed tend to grow?

11. Why does a body weigh less on the equator than at the North Pole? (Give a detailed explanation.)

12. How fast would the earth have to spin for an object on the equator to show zero weight?

13. How does the direction of the centrifugal force change when the revolving body reverses its direction of motion?

14. According to our formula, the centripetal acceleration is $a = v^2/R$; according to another, it is $a = \omega^2 R$. How does the centripetal force change with the radius? Is there a contradiction between these two expressions? Is there anything wrong with this question?

15. Is it physically conceivable for a moon to revolve in a circular orbit about a central planet which is at rest?

16. Is the tension in a pendulum string constant?

17. A string breaks if the mass attached to it reaches the value of 2 kg. How fast must we revolve a mass of 100 gm attached to a length $R = 20$ cm of this string in order to break the string? Find the linear velocity of the mass and the angular velocity of the string.

18. A 3,000-lb car is moving on a curve whose radius of curvature is 60 ft at a speed of 30 ft/sec. Find (*a*) the centripetal force exerted by the road upon the car, (*b*) the centrifugal force exerted by the car upon the road, (*c*) the resultant vector formed by the vector addition of the acceleration of gravity and the centripetal acceleration of the car, (*d*) the force *R* in magnitude and direction exerted by the car upon the road as a result of the combined action of gravity and centrifugal force. How would you design the road in order to avoid slippage at the above speed ("banking")? (Hint: Slippage is due to a force component parallel to the plane of the road.)

19. Find the centripetal acceleration of a cell which is whirled in a centrifuge at the rate of 20,000 rpm 10 cm from the axis of rotation. Assuming the mass of the cell to be 10^{-7} gm, find the centripetal force exerted upon it. How does this force compare with the weight of the cell?

CHAPTER 8

DYNAMICS: 5. ROTATIONAL MOTION

a. Under what conditions will a body continue spinning uniformly about an axis of constant direction?

b. How and in what units do we measure the effectiveness of a force in producing rotation?

c. What kind of motion results from the application of a torque to an unconstrained body?

d. How do we define and in what units do we measure rotational inertia of a body?

e. Does Newton's third law apply to rotational motion?

f. Is there any relationship between angular impulse and angular momentum?

g. What can be said about the angular momentum of an isolated system?

h. Can an isolated body change its angular momentum by changing its moment of inertia?

i. Can it change its angular velocity by changing its moment of inertia?

j. What is a lever?

k. What is the equilibrium condition for a lever acted upon by forces?

l. Can you give examples of devices based on the lever principle?

m. What is the center of gravity of a body?

n. How can it be located experimentally?

o. Under what conditions is a body in equilibrium?

We have discussed in Chap. 7 a necessary condition for the maintenance of rotational motion: the existence of a centripetal force, which is needed to keep a revolving mass point at a constant distance from the center of revolution. But we did not discuss how one can produce rotational motion. A body at rest will remain at rest unless an external unbalanced force is acting upon it. Under what conditions will external forces cause a body to spin?

Rotational Effect of Forces

To answer this question we shall begin with the examination of simple cases from everyday life in which we desire to produce rotational motion of a pivoted object. Figure 8.1 shows the top view of a man trying to open a door by pushing against the handle. Will it be most advantageous for him to apply his force F at A or closer to the hinge at B or C or perhaps at the hinge proper at H? Will it take a lesser force if he applies his force obliquely like f or aiming toward the hinge like P rather than at right angles to the door as in the case of F? We know well from everyday experience that the effort in opening a door depends very much on the point of application of the force as well as on its direction. The same force will be more effective at A than at B or C and entirely ineffective in rotating the door when aimed toward *H*. At any point a force will be most effective when perpendicular to the door and useless when acting parallel to the door toward *H*.

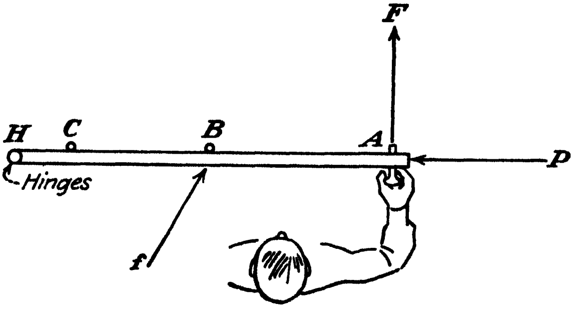


Fig. 8.1. Torque applied to a door.

Moment of a Force

Let us now consider these questions quantitatively, using a familiar device, the lever, as an example (Fig. 8.2).† We shall consider an idealized case of a massless rod loaded with a weight *W* and pivoted at *H* without friction.‡ Forces can be exerted on either side of H to balance the tendency of *W* to rotate the lever. Experience shows that (1) if

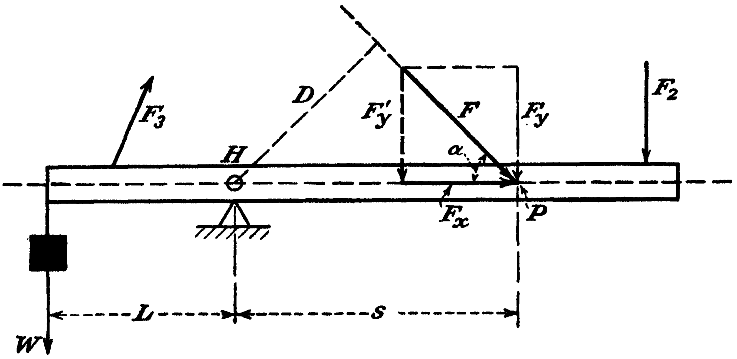


Fig. 8.2. Torque applied to a lever.

---

† In order to have the same effect as *F*, the components of *F* (F_y and F_x) must pass through *P*, the point of application of *F*. The vector $F'_y = F_y$ is part of a figure of graphical computation (illustrating that $\vec{F_x} + \vec{F_y} = \vec{F}$); its position does not imply that F_y does not act on point *P*.

‡ Massless bodies obviously don't exist. Our stipulation simple means that the mass of the lever is chosen so small that its weight is negligible as compared with the applied forces. Such idealizations (neglect of certain properties) are very common in physical reasoning in cases where it is desirable to limit the attention to some essential feature.

an oblique force such as F is resolved into a perpendicular component F_y and one acting parallel to the lever F_x toward (or away from) H, then only the perpendicular component is effective in causing rotation. The magnitude of that component in Fig. 8.2 is

$$F_y = F \sin \alpha \tag{8.1}$$

(2) The point of application P of F_y is of decisive importance: W is maintained in *equilibrium* only if

$$WL = F_y s \tag{8.2}$$

We can modify (8.2) by substituting (8.1) in (8.2). We then get

$$WL = F[(\sin \alpha)s] \tag{8.3}$$

The geometrical meaning of the term in brackets of (8.3) is evident from Fig. 8.2. Since $\sin \alpha = D/s$, $(\sin \alpha)\, s = D$ and we can write

$$\boxed{WL = FD} \tag{8.3a}$$

The meaning of D is as follows: D is the perpendicular distance of the force F (or of its line of action) from the center of rotation H (**fulcrum**). *The product of force and its perpendicular distance from a given point is called* **torque** *or* **moment of the force** *with respect to that point.*

Equation (8.3*a*) *states that a force F can maintain a load W in equilibrium if its moment is equal and opposite to the moment of the load.* We see from this equation that the equilibrium of the lever depends not merely on the applied forces but rather on their moments! **The moment of a force with respect to a point measures its tendency to produce rotation about that point.**

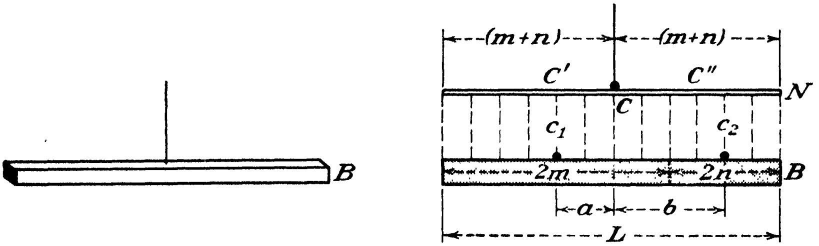
Galileo's Derivation of the Law of the Lever

Galileo has given an ingenious theoretical derivation of the law of the lever. The following is a free rendition of that derivation. He appears to derive the law from a mere consideration of symmetry without recourse to experimentation, but actually the derivation contains hidden assumptions which are suggested by experience.

Consider a uniform, heavy prismatic bar B (Fig. 8.3*a*). At what point should it be suspended in order to remain in equilibrium? We don't hesitate to choose the center point as the point of suspension, since in this symmetrical configuration we see no reason why the bar should tip over to the left rather than to the right side or vice versa. Let us now suspend this bar B from a weightless bar N of equal length $L = 2(m + n)$ (Fig. 8.3*b*) which will serve as a lever. The suspension is accomplished by many massless strings (dotted lines).

Let us now saw the lower heavy bar into two parts of the lengths $2m$ and $2n$. One string would be sufficient to keep each of these parts suspended horizontally; the rest of the strings could be cut without disturbing the equilibrium. Which of the strings should be left uncut? Again for reasons of symmetry, we make the following suggestion: Leave the strings which run to the centers of the bars $2n$ and $2m$ (c_1 and c_2), and cut the others. The two bars are now equivalent to two weights attached to the upper, massless lever at the distances of a and b from the center point C (fulcrum).

As can be seen from the figure, $a = (m + n) - 2m/2 = n$ and $b = (m + n) - 2n/2 = m$. Let us now assume that W is the weight *per unit length* of our heavy bar. Then the weight of the left part is $W_1 = 2mW$ and of the right part $W_2 = 2nW$. In view of these expressions, the following simple relation is seen to hold for the torques T_1 and T_2 with respect to C of the two weights W_1 and W_2 suspended from the points C' and C'', a centimeters and b centimeters off the center of the lever N:



(a) (b)

FIG. 8.3.

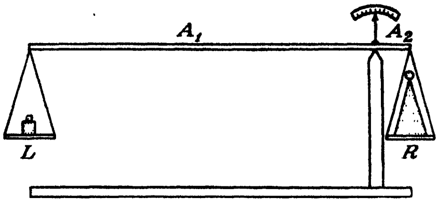
$$T_1 = W_1a = (2mW)n = 2nmW$$
$$T_2 = W_2b = (2nW)m = 2nmW$$

hence, $T_1 = T_2$ and $\boxed{W_1a = W_2b}$

This is **the law of the lever.**

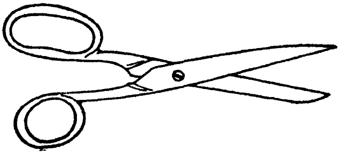
One of the hidden assumptions inherent in this general statement is, for instance, the belief that, if we replace the bars of the weights W_1 and W_2 by spheres or cubes of equal weight, the equilibrium will not be disturbed. Another assumption is that the geographic orientation of the lever is of no importance. Such assumptions are, however, justified by our experience, so that the law has not been deduced a priori, *i.e.*, without recourse to experience.

Figure 8.4 shows a number of devices which utilize the lever principle. The student is to analyze and explain their action in detail.

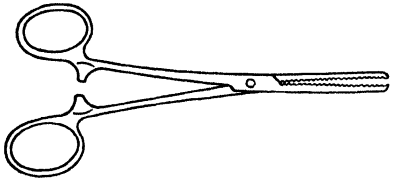


Decimal Scale

Arm A_1 is 10 times as long as arm A_2. The weight on pan L is only $\frac{1}{10}$ of the load on pan R.



Scissors

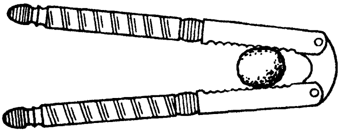


Hemostat

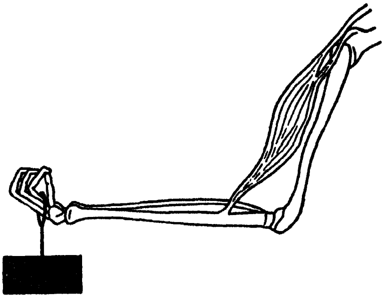
Used by surgeons to clamp a bleeding blood vessel.

Forceps

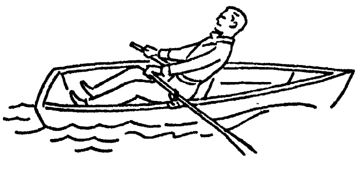
Used by surgeons and by craftsmen to handle small objects.



Nutcracker



Forearm as Lever



Oars

Judo Grip

FIG. 8.4. Various devices utilizing the lever principle. Locate the fulcrum and explain the operation of each device. (*Judo grip illustration from Kuwashima and Welch, "Judo," Prentice-Hall.*)

Accelerated Angular Motion

Application of a torque to a body which is free to rotate about an axis shows that, just as a force is capable of producing linear acceleration of a mass, a moment of force or **torque** is capable of causing angular acceleration of a body. Is there more than a qualitative analogy between the effects of torques and forces? Is there perhaps a law similar to Newton's second law linking the torque, angular acceleration, and inertia of the body?

Suppose we consider that a body which is free to rotate about a fixed axis passing through point A is acted upon by a torque. Let us imagine the body (for simplicity a uniformly flat sheet) subdivided into small compartments (or particles) having the masses m_1, m_2, m_3, ...

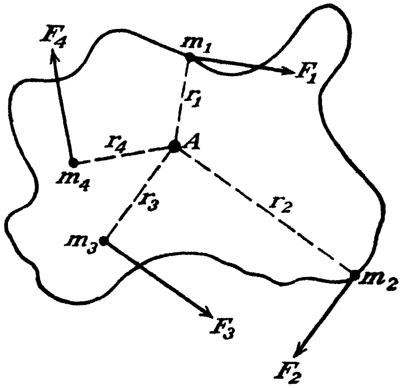


FIG. 8.5. Torque applied to a body rotating about an axis passing through A.

We can imagine the resultant torque to be produced by F_1, F_2, F_3, . . . distributed all over the body, so that each particle is acted on by its share of force.

We are considering only forces perpendicular to the line joining m to A, since the moment of a force or force component, which points toward A, will be zero with respect to A.

We can now apply Newton's second law to the acceleration of our individual particles m_1, m_2, m_3, . . . in the direction of the external forces F_1, F_2, F_3, ...

$$\begin{aligned} F_1 &= m_1a_1 \\ F_2 &= m_2a_2 \\ F_3 &= m_3a_3 \\ &\cdots\cdots \\ F_n &= m_na_n \end{aligned} \tag{8.4}$$

The instantaneous tangential acceleration a of any one of the particles is simply related to the angular acceleration α of the rigid body:

$$a = r\alpha \tag{3.5}$$

(The advantage of introducing the angular acceleration is due to the fact that this quantity has the same value for all points of the rigid body.)

Substituting this in (8.4), we get

$$\begin{aligned} F_1 &= m_1 r_1 \alpha \\ F_2 &= m_2 r_2 \alpha \\ &\cdots\cdots \\ F_n &= m_n r_n \alpha \end{aligned} \tag{8.5}$$

The torque exerted upon each particle is Fr. Hence, we can get an expression for torques T by multiplying both sides of (8.5) by the value of r:

$$\begin{aligned} F_1 r_1 &= m_1 r_1^2 \alpha = T_1 \\ F_2 r_2 &= m_2 r_2^2 \alpha = T_2 \\ &\cdots\cdots\cdots \\ F_n r_n &= m_n r_n^2 \alpha = T_n \end{aligned} \tag{8.6}$$

Adding the preceding equations, we get

$$\Sigma Fr = \alpha \Sigma mr^2 = \Sigma T = T_R \tag{8.6a}$$

where T_R = resultant torque. *This is the desired equation of rotational motion.*

The symbol Σ means sum of. Its meaning is apparent from

$F_1 r_1 + F_2 r_2 + \cdots + F_n r_n = \Sigma Fr$ (read: sum of all products Fr)

Similarly: $m_1 r_1^2 \alpha + m_2 r_2^2 \alpha + \cdots + m_n r_n^2 \alpha$

$$= \alpha(m_1 r_1^2 + m_2 r_2^2 + \cdots + m_n r_n^2) = \alpha \Sigma mr^2$$

Rotational Inertia

Whereas T_R is the resultant torque (sum of all torques), the expression Σmr^2 measures, as we shall see presently, the rotational inertia of a body and has a special name and symbol:

$$\boxed{I = \Sigma mr^2} \tag{8.7}$$

I is the so-called **moment of inertia** of a body. It can be defined as the ratio of the torque to the angular acceleration: $I = T/\alpha$.

Law of Rotational Motion Analogous to Newton's Second Law

The **fundamental law of rotational motion,** which is expressed by equation (8.6*a*), can now be written in terms of new symbols.

$$\boxed{T_R = I\alpha} \tag{8.8}$$

In words: Unbalanced torque equals the moment of inertia times angular acceleration.

Comparing with $F = ma$

we see that in the rotational law torque takes the place of the force, angular acceleration the place of linear acceleration, and the moment of inertia replaces mass.

Since $\alpha = T_R/I$, we see that I acts as "resistance to angular acceleration" much as m acts as "resistance to linear acceleration"; *i.e.*, an increase in I diminishes α.

Can two bodies have the same mass and yet different moments of inertia? The answer is yes. The moment of inertia depends not only on mass but equally on its distribution. For instance, for a *small* sphere swung by a string, $I \simeq mr^2$ (where r is the distance from the center of the axis of rotation to the center of the sphere, the radius of the sphere being negligible as compared with r). If we double the length of the string, we quadruple the moment of inertia without changing the mass. Or to name another example, the moment of inertia of a disk is smaller than that of a torus (doughnut) of equal mass, since in the latter case Σmr^2 is larger.

Law of Rotational Motion Analogous to Newton's Third Law

The rotational analogue to Newton's third law is as follows: **If body *A* exerts a torque upon body *B* with respect to a given axis, then body *B* exerts an equal and opposite torque upon *A* with respect to the same axis.** (The derivation of this statement from Newton's third law is quite simple and is left to the student.)

Law of Rotational Motion Analogous to Newton's First Law

The rotational analogue of Newton's first law follows from (8.8) by placing $T_R = 0$:

A body (of constant moment of inertia) will remain in uniform rotation about an axis unless acted upon by an external unbalanced torque.

Angular Impulse and Angular Momentum

The analogy to the linear case can be extended further to the derivation of the law of conservation of angular momentum, which has very interesting implications and applications:

From (3.2), $$\alpha = \frac{\omega_2 - \omega_1}{t}$$

Substituting this in (8.8), we get

$$T_R = I\frac{\omega_2 - \omega_1}{t}$$

hence, $$\boxed{T_R t = I\omega_2 - I\omega_1} \tag{8.9}$$

This equation is analogous to

$$Ft = mv_2 - mv_1 \tag{5.2}$$

We call the product Tt of torque and time **angular impulse** and $I\omega$ the **angular momentum.** Equation (8.9) states: **The change in angular momentum of a body is equal to the angular impulse which caused this change.**

Conservation of Angular Momentum. This theorem may also be applied to a system of bodies, in which case the left side signifies the resultant angular impulse of all the external torques acting upon the system and the right side stands for the vector sum of all changes in angular momentum which have occurred within the system. (A vector may be used to represent an angular momentum much in the same way as it can be employed to symbolize a torque.)

Applying equation (8.9) to an isolated system of bodies, we obtain†

$$\underline{\Sigma Tt} = \Sigma(I\omega_2 - I\omega_1) = \underline{\Sigma\Delta(I\omega)} \qquad (8.9a)$$

The implications of (8.9a) in the case of $\underline{Tt = 0}$ (no external torques acting upon the system) are particularly interesting and important. In this case

$$\boxed{\Sigma\Delta(I\omega) = 0} \qquad (8.10)$$

(In words: The sum of all changes in the angular momentum is zero.) **This is the law of conservation of angular momentum:**

The angular momentum of an isolated system which is not subjected to the action of external torques remains constant (in magnitude and direction).

We shall now examine some physical consequences of equation (8.10) as an example of how to interpret the physical meaning of a mathematical statement.

Examples. Let us consider a simple system consisting of two bodies: you and the earth. Suppose the earth does not rotate to begin with. If you start running clockwise about the equator, you will acquire a certain angular momentum with respect to the axis of the earth. The angular momentum of this two-body system was zero to begin with in the absence of rotation. According to equation (8.10), it should remain zero, since no torques are acting from the outside! For the resultant angular momentum ($I\omega$) of the system to remain zero, the entire earth must acquire an angular momentum equal and opposite to yours! Since the earth actually does rotate to begin with, your equatorial trip around the world will either slow down or speed up the globe according to whether you walk in the direction of rotation or in the opposite sense.

---

† Compare with equation (5.2).

The following example provides an additional illustration: Gun barrels are usually rifled so that a spin is imparted to the bullet as it leaves the muzzle. Thus in addition to the backward recoil, which is a consequence of the law of conservation of linear momentum (Chap. 5), the rifle suffers a slight "angular recoil." The latter is a consequence of the law of conservation of angular momentum; the rifle must acquire an angular momentum equal and opposite to that which was imparted to the bullet inside the muzzle.

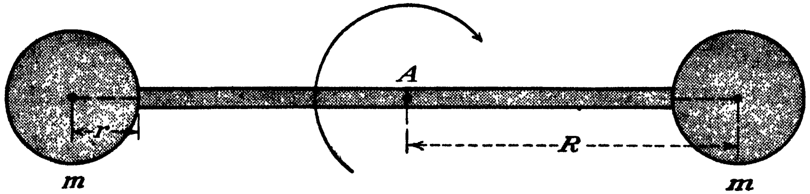


FIG. 8.6. Dumbbell having a moment of inertia $I \simeq 2mR^2$.

Let us now consider a quantitative example: Fig. 8.6 shows an isolated body in the shape of a dumbbell rotating about an axis passing through A. If R is very large compared with r, its moment of inertia about point A is $I = 2mR^2$ (approximately). What will happen to the angular velocity ω if the radius R is reduced to 50 per cent of its original size, thus changing the moment of inertia of the body? According to equation (8.10), the final $I\omega$ should be equal to the original value:

$$\Delta I\omega = I_1\omega_1 - I_2\omega_2 = 0$$

hence,
$$I_1\omega_1 = I_2\omega_2 \qquad \text{or} \qquad \underline{\frac{\omega_2}{\omega_1} = \frac{I_1}{I_2}}$$

Now, $I_1 = 2mR^2$. To get I_2 we must replace R by $R/2$.

Then
$$I_2 = 2m\left(\frac{R}{2}\right)^2 = 2m\frac{R^2}{4} = \frac{m}{2}R^2$$

hence,
$$\frac{\omega_2}{\omega_1} = \frac{I_1}{I_2} = \frac{2mR^2}{\frac{1}{2}mR^2} = \frac{4}{1}$$

that is, *ω is increased fourfold by halving R.*

This result is very interesting. Whereas according to the law $F = ma$ a body cannot acquire a linear acceleration in the absence of an external force, the law $T = I\alpha$ does not exclude the possibility of a body's acquiring an angular acceleration in the absence of an external torque, provided the moment of inertia of the body can be varied by forces applied from within the body.

A jumping diver uses this mechanism for effect as well as to control his "launching." When he leaves the diving board, he imparts to himself an angular impulse. As long as his body is stretched out, its moment of inertia is high and he spins slowly. By curling up he can reduce his moment of inertia and thus speed up his rotation. By controlling the curvature of his body, he can control his angular velocity.

The same principle can also be used to demonstrate convincingly the rotation of the earth. Although the demonstration would also work in the latitude of Chicago, we picture, for simplicity, the experiment as being performed at the North Pole. Imagine a long, horizontal bar carrying two heavy masses at its ends and shaped and suspended as shown in Fig. 8.7. The spheres m_1 and m_2 are prevented by an appropriate mechanism from falling down toward the center. The whole structure is brought to rest with respect to the earth. This, however, means that it is sharing the angular velocity ω_1 of the earth; *i.e.*, it spins about the celestial north-south axis once around in a sidereal day just like the earth. Now, suddenly, m_1 and m_2 are freed to roll down toward the center. This reduces the moment of inertia of our "horizontal pendulum," which has the original angular velocity ω_1 with respect to the stars. According to the law of conservation of angular momentum (8.10), we should expect ω to increase now to a new value $\omega_2 > \omega_1$. This means that our "torsion pendulum" must suddenly rotate about its axis of suspension, getting ahead of the earth. And it actually does!

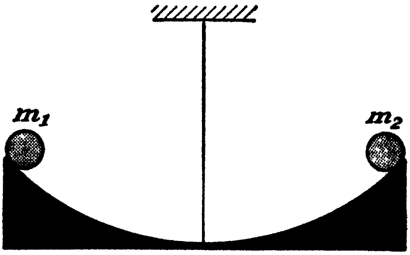


FIG. 8.7. Apparatus for the demonstration of the rotation of the earth.

The possibility of controlling the angular velocity of a body by changing its moment of inertia suggests the alarming possibility that we may prolong the day by slowing the rotation of the earth every time we raise our hand, thus increasing the moment of inertia of the globe. Minor periodic variations of the length of the day have actually been observed, but they have a different cause. They are attributed to slight periodic deformations of the terrestrial globe, which affect its moment of inertia. (Try to estimate the order of magnitude of the effect of raising a hand upon the length of the day.)

APPENDIX

Equilibrium of Bodies and the Center of Gravity

We have seen in the appendix of Chap. 7 that a point is in equilibrium when the vector sum of all forces acting upon it is zero. Is the same true of an extended body?

Let us consider the following simple experiment. We take a negligibly thin, heavy, uniform rod which weighs 2 lb and place its center on a sharp edge *E* which is placed on a spring scale (Fig. 8.8). The reading of the scale is now 2 lb, and our rod is in equilibrium under the action of a 2-lb force exerted upward by the scale and a 2-lb force of gravity exerted by the earth. This downward force of gravity *W* is obviously the resultant of the gravitational forces *F* which act on the individual portions ("elements") of the rod into which we can imagine the rod to be subdivided.

We know from experience that, if we were to place the end *A* or *B* of the rod on the edge *E*, it would not remain in equilibrium but would rather turn over clockwise or counterclockwise, respectively.

It is easy to see from the diagram that the forces *F* possess a moment with respect to the edge *E*. It is only in the symmetrical position illustrated in Fig. 8.8 that the sum of the clockwise moments of the forces to the right of edge *E* is equal to the opposing moment of the forces to the left of *E*. Under these conditions the sum of all the moments due to the weights *F* of the elements of the rod is zero, and the rod remains in equilibrium, but if we were to shift the point of support *E* toward *A* or toward *B*, the rod would turn over clockwise or counterclockwise, respectively. **The point with respect to which the sum of all gravitational torques is zero is called the** *center of gravity.* It is the point at which the equilibrating force which neutralizes gravity must attack in order to prevent the body from moving and rotating owing to the action of gravity. Since the equilibrating force and the resultant *W* of all the gravitational forces which act on the body are acting at the same point as "action" and "reaction," we can also say that **the center of gravity is the point through which passes the resultant force of gravity acting on the body. In all problems of statics, the entire weight of an extended body may be imagined as being concentrated at the center of gravity.**

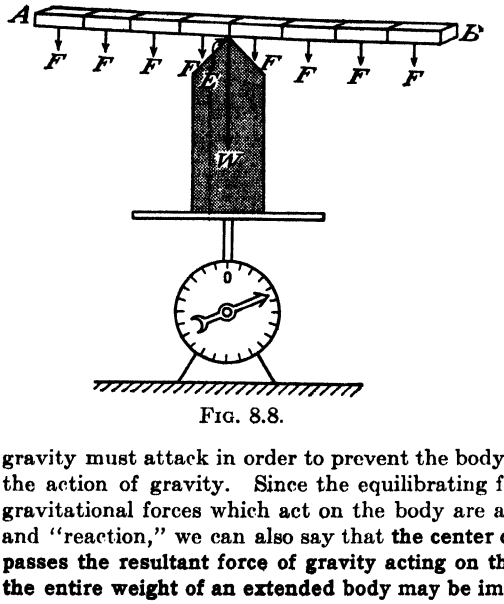


FIG. 8.8.

If we support an extended body at an arbitrary point *P*, the body tends to turn until the center of gravity *C* comes to lie under the point of support which acts as a fulcrum (Figs. 8.9*a* and *b*). We see from Fig. 8.9*a*, where the entire weight *W* is shown as acting at point *C*, that a counterclockwise torque rotates the body until this torque vanishes when *C* comes to lie under point *P* (Fig. 8.9*b*).

Suppose our body is a piece of sheet metal. If we suspend it at point *P* and draw the vertical line *AB*, we know the center of gravity lies on this line. If we now suspend the body at a different point *P′* (Fig. 8.9*c*) and draw the vertical line *DE*, we know that the center of gravity must lie somewhere on this line and hence at the intersection of the lines *AB* and *DE*. If we now pierce our sheet metal at the point *C* and put a needle through the hole, we can rotate the body as we please and it will always remain in equilibrium.

An extended body is in equilibrium when the following two conditions are fulfilled:

1. **The sum of the forces acting on the body must be zero.**
2. **The sum of the torques acting on the body must be zero.**

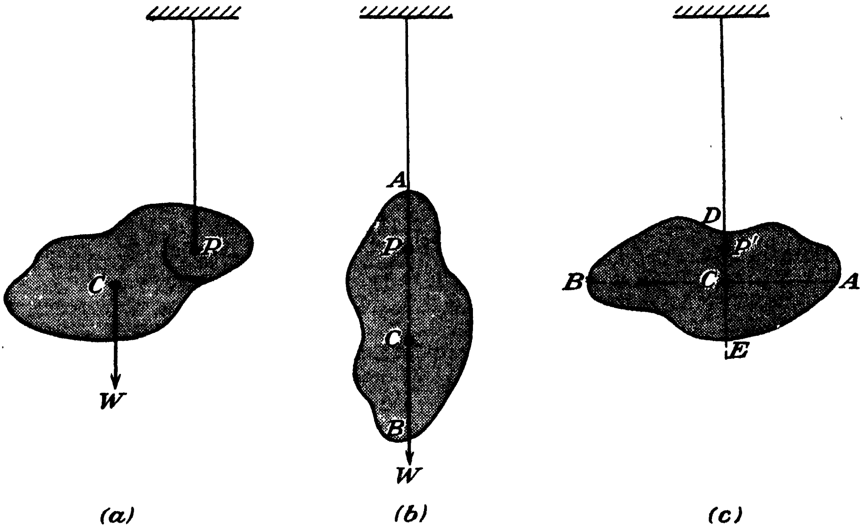
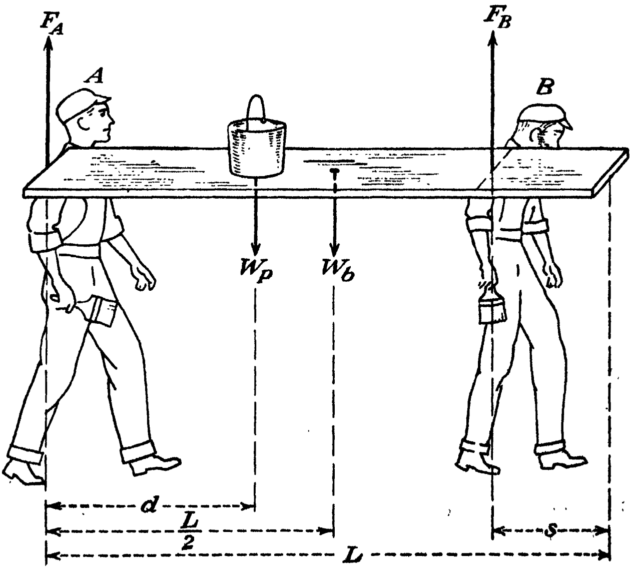


Fig. 8.9.

Example

Two painters carry a "walkboard" on top of which stands a pail of paint (see figure). The weight of the walkboard is $W_b = 24$ lb. Its length is $L = 8$ ft. The weight of the pail of paint is $W_p = 20$ lb. The center of the pail is $d = 3$ ft from the left end of the walkboard. The left man supports the walkboard at its left end, whereas the right man rests the walkboard on his shoulder, $s = 2$ ft away from its right end. How much weight does each man carry?



Solution. Consider the walkboard as a body which is maintained in equilibrium by the following forces which act on it: (1) the force of gravity W_b attacking at the center of gravity which lies at the center of symmetry of the board, (2) the weight of the pail W_p, (3) the reaction F_A at the shoulder of the left man, and (4) the reaction F_B at the shoulder of the right man. According to the first condition of equilibrium, the vector sum of all these forces must be equal to zero. Taking forces pointing upward arbitrarily as positive, we get

$$F_A + F_B - W_p - W_b = 0 \tag{1}$$

substituting given data, we obtain

$$F_A + F_B - 20 - 24 = 0; \qquad F_A + F_B = 44 \text{ lb} \tag{1a}$$

According to the second condition of equilibrium, the sum of the moments with respect to any arbitrarily chosen point must be zero. We choose, for convenience, the edge of the left end of the walkboard as our reference axis and take counterclockwise moments with the positive sign, clockwise moments with the negative sign. We read off from our diagram

Force, lb	Lever arm, ft	Torque, lb-ft
F_A	0	$F_A \cdot 0 = 0$
$W_p = 20$	$d = 3$	$-W_p \cdot d = -20 \cdot 3 = -60$
$W_b = 24$	$\frac{L}{2} = 4$	$-W_b \frac{L}{2} = -24 \cdot 4 = -96$
F_B	$L - s = 8 - 2 = 6$	$+F_B(L - s) = +6F_B$

Hence, we obtain the following equilibrium equation:

$$\text{Sum of the torques} = -60 - 96 + 6F_B = 0 \tag{2}$$

From this follows: $6F_B = 156$; $\underline{F_B} = \frac{156}{6} = \underline{26 \text{ lb}}$

Substituting this value into equation (1*a*), we obtain F_A, the weight carried by the other man:

$$\underline{F_A} = 44 - 26 = \underline{18 \text{ lb}}$$

QUESTIONS AND PROBLEMS

1. Find the moment of inertia of a hoop of radius 40 cm weighing 500 gm.

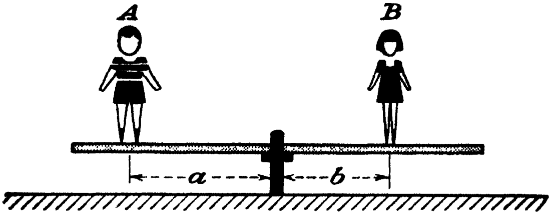
2. A torque is applied for 2 sec to the hoop of the preceding problem. As a result the hoop acquires the angular speed of 2 rad/sec.

a. How large is the torque?

b. How large is the angular momentum acquired?

3. The angular speed of a wheel changes from 3 to 9 rad/sec in 2 sec when a torque of 10 lb-ft is applied to it. Find the moment of inertia of the wheel.

4. A merry-go-round is improvised by mounting a wide board of negligible mass on a practically frictionless axis. Two children are standing on the board, as shown in the figure, being free to choose and change their positions. Their weights are $W_A = 64$ lb and $W_B = 32$ lb. The board is 10 ft long. The initial position of the children is as follows: $a = 4$ ft and $b = 3$ ft.



a. Somebody applies a force of 10 lb for 2 sec to one end of the board at right angles to the board. How large an angular speed does the merry-go-round acquire?

b. How does this angular speed change when the children walk to the ends of the board and stay there?

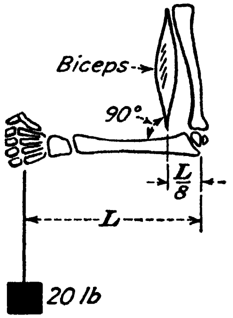
c. What happens to the angular speed when the children get both on the same side of the bar and stand midway between its end and the axis?

5. A bicycle wheel of mass = 1 kg and of R = 40 cm rotates at the rate of 2 revolutions per second. A frictional torque of $T = 10^7$ dyne-cm is applied to it for 4 sec. Find the final angular velocity of the wheel. (Assume the mass of the wheel to be concentrated at its rim.)

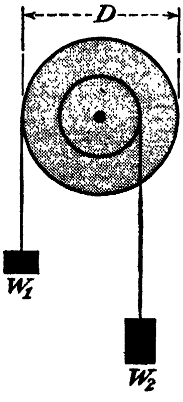
6. An aluminum plate of moment of inertia I is floating on the surface of mercury. You drop a spinning aluminum plate (angular velocity ω_1) whose moment of inertia is $3I$ concentrically on the floating plate. With what angular velocity do the two plates continue spinning? (Neglect fluid friction.)

★7. If you were a cat falling with all your four legs pointing upward, what would you do in order to land on four paws? (Describe the motions in detail.) Is such a turn possible according to the law of conservation of angular momentum?

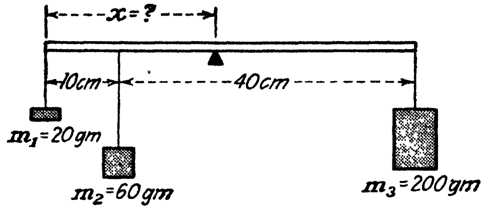
8. How large a force does your biceps muscle exert when you lift a 20-lb weight by flexing your elbow (see figure)?



9. The double pulley shown in the figure is in equilibrium. The weights are W_1 = 3 gm, W_2 = 12 gm. The diameter of the larger pulley is D = 10 cm. Find the diameter of the smaller pulley.



10. Where would you support this bar so that it would remain in equilibrium?



11. Two men are carrying an 8-ft-long uniform ladder which weighs 30 lb. One of the men holds the ladder at one end and the other man holds it 3 ft from the other end. How much weight does each man carry?

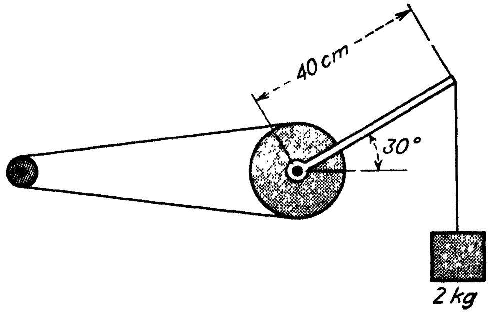
12. If you stand with your back against a wall, keeping your knees unbent, your heels touching the wall, you will find it impossible to bend foreward to pick up an object from the floor without falling forward. How is this to be explained?

13. A horizontal bar is suspended by means of two ropes. One rope is tied to one end of the bar while the other rope is tied 3 ft away from the other end. The bar is 10 ft long. The tension in the rope which is tied to the end of the bar is 5 lb.

a. How heavy is the bar?

b. How large is the tension in the other rope?

14. How large is the torque exerted by the force shown in this diagram?



15. If the same torque is applied for the same time to a sphere and to a cylinder of equal mass and density whose height is equal to half the radius of the sphere, which acquires the greater angular speed?

CHAPTER 9

WORK AND ENERGY: 1

a. Can a force cause acceleration? Can work cause acceleration? Can an impulse cause acceleration?

b. Can we ever detect an unbalanced force unless it exerts an impulse or does work?

c. What is the distinction between force and work?

d. What is the distinction between work and energy?

e. How much work does the earth do on the moon when the moon revolves around the earth? (Assume a circular orbit.)

f. What is the physical meaning (1) of negative work? (2) of negative kinetic energy?

g. Can you name some forms of energy?

h. If two stars were to fall into each other as a result of mutual attraction, they would collide with considerable kinetic energy. Where does this energy come from?

i. A stone falls to the ground from a height of 10 ft. How does its energy just before it hits the ground compare with its initial energy content?

j. When you throw a ball upward, it stops at a certain height. What happens to its kinetic energy?

k. How could you detect the potential energy of a body?

l. A boy on a train speeding at 60 mi/hr holds a 1-lb apple. What is the kinetic energy of the apple? What answer might the boy give to this question?

m. How much work is done on a weight in moving it once around a vertical circle of radius R?

n. What is an equipotential surface?

o. What is the relationship between equipotential surfaces and the direction of the gravitational force?

p. When two elastic balls collide, they stop temporarily. What happens to their kinetic energy? What happens to the shape of the balls?

q. Can you store energy by doing work against friction?

r. In what kind of processes may energy periodically change form?

s. How do you define the mechanical advantage of a machine?

t. How do you define the efficiency of a machine?

Force, Work, and Energy

A force is an agent which is capable of neutralizing gravity. For instance, in Fig. 9.1 weight A keeps a body B in equilibrium by exerting upon it an upward force equal to the weight of B.

The ability of A to exert this force on B is apparently inexhaustible.

If we now place a tiny overweight on A, it will descend nearly uniformly (actually with a negligible acceleration), raising B against the force of gravity. *The body A is said to be doing work on body B by raising it.* The ability of A to do work on B is, however, not unlimited. This process will stop as soon as A reaches the floor. This exhaustible something that enabled A to keep lifting B against gravity as long as A was above ground and that it seems to lose as it reaches the floor we call **energy** (from the Greek, ability to do work).

The terms *force* and *work* are frequently used interchangeably and vaguely by nonphysicists. Even in physics this confusion persisted for a long time. Leonardo da Vinci was perhaps the first to have made a clear distinction between force and work. He remarked that work involved motion of a body in the direction of the moving force. But his contributions remained long unpublished and thus had no influence on the development of scientific concepts.

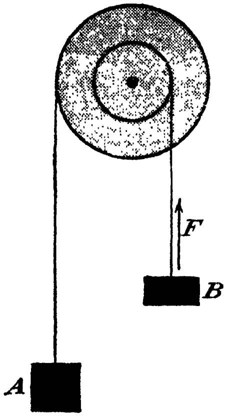


Fig. 9.1.

Definition of Work. Work done on a body by a given force F is defined as the product of the applied force and the displacement of the body in the direction of the force.

Or the following is equivalent to the above statement:

The work done on a body by a given force F is defined as the product of the displacement of the body times the component of the force in the direction of this displacement. (The units of work are units of force times units of distance, for instance, foot-pounds or dyne-centimeters.)

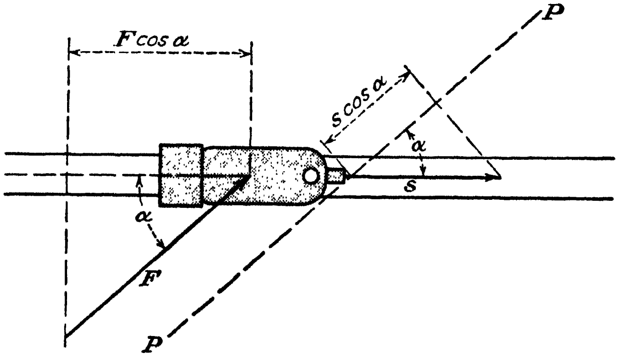


Fig. 9.2. Work done by a force which is not parallel to the displacement.

The equivalence of these two definitions can be easily seen from Fig. 9.2. A locomotive constrained to move on rails is equipped with a sail. The wind exerts a constant force F upon the sail and causes the locomotive to move to the right through a distance s. How much work did the force of the wind do on the locomotive?

First Version. Component of the force in the direction of displacement is $F \cos \alpha$.

Hence, $$\underline{Work = (F \cos \alpha)s} \tag{9.1a}$$

Second Version. To find the component of the displacement in the direction of F, draw a line PP parallel to F as shown in Fig. 9.2. The projection of s upon the line PP (*i.e.*, the component of s in direction of the force) is, as can be seen from the figure, $s \cos \alpha$.

Hence, $$\text{Work} = F(s \cos \alpha) \tag{9.1b}$$

Comparing equations (9.1a) and (9.1b), we see that they are identical.

What Is Energy?

The lifting of a weight in Fig. 9.1 is a particularly simple case of doing work, since the force and displacement have the same direction. We can use this simple performance as a "reagent" for the energy content of a body. *A body will be said to possess* **energy** *if it is capable of lifting a weight.* In general, a body is said to possess energy if, and only if, it can do work.

Kinetic Energy

In certain cases the ability of a body to do work is quite apparent. For instance, if we watch a moving object, we recognize immediately that it possesses energy due to its motion relative to us. We could, for instance, make it do work by lifting its own mass against the force of gravity by allowing it to roll up an inclined plane (Fig. 9.3). The ball will eventually stop at the height B. At this point its velocity has vanished and its ability to lift its weight against gravity is exhausted. This is an example of the so-called **kinetic energy** of a body which enables it to do work by virtue of its motion. This example shows how kinetic energy becomes exhausted as more and more work is performed at its expense.

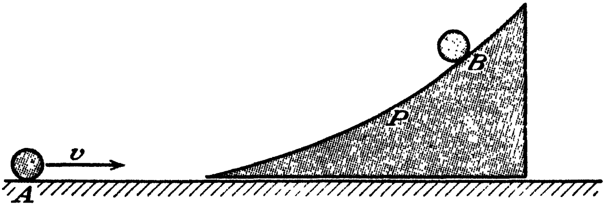


FIG. 9.3. Conversion of kinetic energy into potential energy.

The Relative Character of Kinetic Energy. It is also important to note the relative character of the kinetic energy. If the inclined path P

of Fig. 9.3 were moving to the right with the same velocity as our sphere, the ball would be chasing it without ever being able to rise up on it. With respect to an observer moving with the inclined path P, the ball is at rest and has no kinetic energy. It is only the relative velocity between the ball and the plane P which determines how high the ball will rise. You can, for instance, make a ball which is initially at rest rise by moving the path P toward it with the velocity v.

Potential Energy

Let us now examine what happens to the original kinetic energy of the ball of Fig. 9.3 when it reaches its highest point and stops. The kinetic energy is evidently gone! But not for long, because the ball now starts rolling down and arrives at the bottom, moving in the opposite direction with nearly its original velocity. The sphere has been, so to speak, reflected backward. As we perfect the smoothness of the surfaces of the sphere and of the inclined path, the final velocity of the "reflected" sphere is more and more nearly equal to its original velocity and so we can extrapolate from our experience to the ideal case of perfect smoothness, in which case the original kinetic energy should be fully restored.

Even though in the above example we see kinetic energy disappear temporarily, we are disinclined to say that energy has vanished, because we see it reappear again. *Our observations permit us to ascribe to energy the property of indestructibility and to treat it as if it were a substance which can masquerade in different equivalent disguises but cannot disappear wholly or partly.* The concept of an indestructible energy proves of great practical value and convenience. In terms of this concept, one would describe the above observations by saying that, when the sphere comes to rest, its original kinetic energy has been transformed into a different form of energy, **potential energy**, or energy of position. The elevated position of the ball enables it now to do work by descending down the inclined path. The available work for the sphere shown in Fig. 9.4 is given by the product of L (the available length of the path down the incline) times the component of W in the direction of L, $W \sin \alpha$. Hence, loss of potential energy $=$ work done $= (W \sin \alpha)L = W(L \sin \alpha)$.

We recognize from Fig. 9.4 that $L \sin \alpha = h$ is the vertical displacement of the ball; hence, we can write

Loss of potential energy: $$\boxed{PE = Wh} \qquad (9.2)$$

The significant feature of this equation is the disappearance of the angle α. *The change in potential energy depends evidently only on the* *vertical displacement and is independent of the inclination of the path of the body.* In fact, even if the inclination of the plane should constantly vary, *i.e.*, if it be a fancy curve, the kinetic energy acquired by the ball at the bottom of its path will be independent of the shape of the curve, being determined only by h, the height of descent. (The independence of the kinetic energy of the shape of the path can be made plausible by considering the curved path as being made up of many, very short, straight inclined paths of varying inclination.)

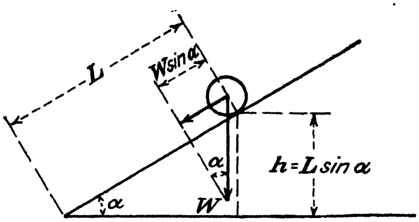


FIG. 9.4.

It is left to the student as an exercise to explain why the work expenditure (and hence change in potential energy) in lifting a weight along the different paths shown in Fig. 9.5 is the same. (Hint: Work = $Fs \cos \alpha$ is positive as long as the angle between the force and the displacement is smaller than 90° and negative for $\alpha > 90°$, since then $\cos \alpha$ is negative. Approximate the curved paths by a sequence of short straight lines. Resolve each of these linear displacements into a vertical and a horizontal component. Remember that no work is done in a horizontal displacement and that for upward displacements the work of the force of gravity is negative whereas it is taken with the positive sign for downward displacements.)

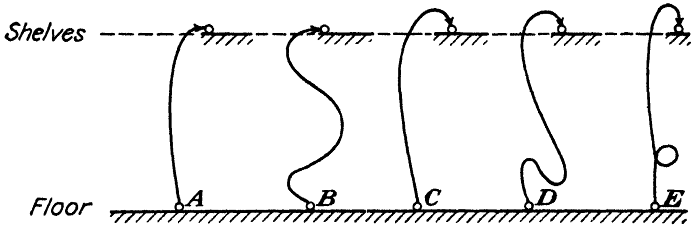


FIG. 9.5.

Conversion of Potential into Kinetic Energy. To determine the quantitative relationship between the loss of potential energy and the gain in kinetic energy, let us imagine a mass m acted on by a force F which may, for instance, be the force of gravity. The mass will be accelerated and will lose potential energy as it falls. According to equation (9.2), the loss is PE = Fs, where s is the distance the body falls. According to the idea of conservation of energy, the amount of kinetic energy gained by the body is to be considered as the equivalent of the lost potential energy. How large is the gain in kinetic energy which is equivalent to the potential-energy loss Fs?

To answer this question we shall attempt to express the potential energy loss in terms of the gain in velocity: Loss of potential energy (PE) = (gravitational) work done on the accelerated body = gain in kinetic energy (KE):

$$\text{PE} = Fs = \text{KE} \qquad (a)$$

Since the motion under the influence of the constant force F is uniformly accelerated, the displacement s is given by

$$s = \frac{a}{2}t^2 \qquad (b)$$

and F is given by

$$F = ma \qquad (c)$$

(c) and (b) in (a):

$$\text{KE} = (ma)\left(\frac{a}{2}t^2\right) = \frac{m}{2}(at)^2 \qquad (d)$$

But for uniformly accelerated motion, $v_f = v_0 + at$, and if $v_0 = 0$,

$$v_f = v = at \qquad (e)$$

Substituting (e) in (d), we get

$$\boxed{\text{KE} = \frac{m}{2}v^2 = Fs} \qquad (f)$$

This equation expresses the kinetic energy of a body in terms of its velocity and mass regardless of how this energy was gained.

If we know the velocity of a body, we can easily compute its **KE as the half product of its mass and the square of its velocity.**

Relative Character of Potential Energy. We saw that work against gravity is associated with an increase in potential energy and that the potential energy of an object is determined by its weight and position. The potential energy of a body is measured by the work which it can do by descending, (Wh). But where do we measure h from? This question calls to our attention the relative nature of potential energy. It has no meaning to speak of potential energy of a body unless we specify with respect to what level it is measured. The level of reference can be chosen arbitrarily.

Storage of Potential Energy

Any surface to which the pull of gravity is perpendicular at every point is a so-called **equipotential surface,** *i.e.*, surface of constant potential energy. If our earth were strictly a sphere, the equipotential surfaces would be spherical surfaces. The surface of an ocean engulfing the globe would be a spherical equipotential surface. No work would be required to displace an object along this surface, since the tangential displacement would be perpendicular to the radial force of gravity. Any change in gravitational potential energy of a body is evidently connected with the process of moving the body against the force of gravity. Can we similarly store potential energy in a body by moving it against other than gravitational forces? The answer is yes. We shall become acquainted later with electrical and magnetic forces which show a similar behavior. The following is an example of the so-called *elastic* forces and elastic potential energy. When you compress a spring, you are doing work against an elastic restoring force. As a result you accumulate potential energy in the compressed spring which is capable of doing work on a body as it expands. Elastic springs are used very commonly to store potential energy in devices such as the pocket watch or the kymograph.

Frictional Forces. Can we store potential energy by doing work against any kind of forces? No. For instance, if you are pushing a box along the floor, you are doing work against frictional forces, but when you release the box, it does not spring back with kinetic energy equivalent to the work you spent in pushing it forward. **We shall define as frictional forces, forces which are of such a nature that work done against them does not result in storage of mechanical energy.**

Since the frictional force always opposes the direction of motion of a body, it acts as a retarding force and dissipates kinetic energy. What has happened to the work we did on the box in the above example in moving it? Has that energy been lost? We shall answer this question in Chap. 10. For the time being, we shall merely note that, as far as we can see, the work which we have done on the box has not given rise to kinetic energy.

Oscillatory Transformations of Energy

Is an interconversion possible between elastic potential energy and kinetic energy similar to what we have seen take place in the case of an inclined plane between the kinetic and gravitational potential energy? Such an interconversion is possible and is exemplified by a bouncing ball. As a moving ball strikes a horizontal wall, work is done against elastic restoring forces in compressing the ball. When the ball stops, it is in a state of maximum elastic deformation. Elastic potential energy has been stored in it. Now the ball begins to expand. It is losing elastic energy, but at the same time it gains kinetic energy, and by the time it has assumed its original spherical shape, it is moving with its original velocity in reverse. It has been reflected from the wall. (This description assumed a case of ideal elasticity, *i.e.*, perfect resumption of the original shape.)

A perfectly elastic ball which bounces repeatedly up and down illustrates a periodic interconversion between kinetic and potential energy. Regular periodic conversion of energy from one form into another and back again is of fundamental theoretical and practical importance in physics. A classical example of such a process is a swinging pendulum. In its position of maximum deviation from the equilibrium position, the pendulum bob is also at its highest point. It is at rest and has gained potential energy. When the pendulum bob reaches the mid-point, it is in its lowest position, and hence its potential energy has its minimum value, which we can arbitrarily call zero if we take the lowest position of the pendulum bob as our zero level. Its velocity is, however, at a maximum at that moment, and all its mechanical energy is now present in the form of kinetic energy. Inertia carries it past its equilibrium position until it reaches the opposite maximum displacement, at which point all its kinetic energy has been again transformed into potential energy. At any intermediate point the energy is present partly in kinetic and partly in potential form and at all times (neglecting friction) **the sum of potential plus kinetic energy remains constant.** Actually, however, owing to friction the mechanical energy of the system diminishes from swing to swing. The excursions of the pendulum become smaller, until it stops. Such an oscillation is called a **damped oscillation.**

A similar process can be initiated by disturbing the equilibrium of a weight suspended on a spring. It will go up and down while three forms of mechanical energy will play a part in the periodic conversion process: (1) kinetic energy of the weight (neglecting the mass of the spring), (2) gravitational potential energy of the weight, (3) elastic energy of the spring. When the mass is *at the highest point,* (*a*) KE = 0, (*b*) PE is at a maximum, (*c*) elastic energy of the distended spring is at a minimum. When it is *at the lowest point,* (*a*) KE = 0, (*b*) PE is at a minimum, (*c*) elastic energy of the distended spring is at a maximum.

Generalized Concept of Oscillation. We shall see later that there are other forms of energy besides the mechanical ones. They are also capable of undergoing transformations from one form into another. Thus, we can tentatively generalize the concept of oscillation by extending it to all forms of energy, even to cases where no vibratory process is visible. An outstanding example of such a process will be the electromagnetic oscillations, which will be studied in Chap. 33. **We shall define as oscillation any periodic interconversion between two forms of energy.**

Simple Machines

When considering transformations from one form of mechanical energy into another, we should not omit the simplest case, namely, the transformation of potential energy into potential energy. This may look at first like a misprint. What would be the sense of such a "conversion"? Let us look at a practical example: Suppose you want to raise an object weighing 300 lb (and exceeding your lifting ability) off the ground. You could accomplish it very easily by improvising a lever (Fig. 9.6). If you choose, for instance, a lever arm ratio of 2:1, then stepping on the lever at the point A will enable you by applying your body weight of 150 lb to lift the load of 300 lb located at point B [see equation (8.3a)].

As W_B is lifted through the vertical distance h_B and thus made to gain potential energy, your body, standing at the point A of the lever, sinks through the vertical distance h_A, losing potential energy. The potential energy which you had in the initial position of the point A has been transferred to the weight at B.

Since the arm ratio is 2:1, the ratio of the vertical distances traversed by W_A and W_B is also 2:1. Our body weight moved twice as far as W_B, and hence, energy has been neither lost nor gained by this system: $W_A h_A = W_B h_B$. We have not done less work in lifting W with the aid of the lever than we would have done in lifting it directly. There is, however, *a distinct practical gain in using the lever due to the fact that we have been able to overcome a large force W_B by applying a smaller force W_A. The price for this advantage was paid in terms of the larger distance the smaller force had to move.*

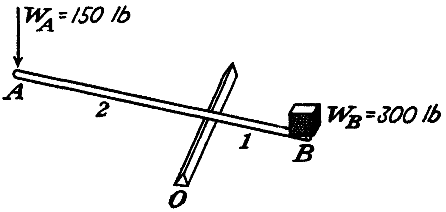


FIG. 9.6. The lever as a simple machine.

The lever illustrates a "simple machine," *i.e., a device which makes it possible to do work by means of a larger force "put out" by the machine than one applies to the machine "input."*

Conservation of Work. Experiments with all types of machines lead to the following result (for the limiting case of zero friction):

$$F_1 s_1 = F_2 s_2 \tag{9.3}$$

(F_1 = force input, F_2 = output force, s_1 and s_2 = distances covered by the points of application of input force and output force in the direction of the force.)

This equation states: Work input is equal to work output. Thus, the use of a frictionless (ideal) machine involves neither a loss not a gain in energy. This **law of conservation of work** is a special case of a general **law of conservation of energy** (see page 115) which was induced from many restricted generalizations of experience such as the preceding one.

Simple machines (such as, for instance, the lever) **are designed to help overcome large forces by exertion of smaller ones and to enable one to apply the force in a conveniently chosen direction.** The ratio of $\frac{\text{output force}}{\text{input force}}$ is called the actual **mechanical advantage** of a machine. Simple machines such as pulleys, levers, screws, hydraulic presses, and other elements are commonly employed in the mechanical design of household tools and of industrial and medical equipment. Special problems will serve to familiarize the student with some of these devices.

The efficiency e of a machine is defined as follows:

$$e = \frac{\text{work output}}{\text{work input}} \tag{9.4}$$

In the ideal case (no friction) $e = 1 = 100$ per cent.

Actually, however, the work input serves not only to deliver the desired work output but also to overcome inevitable frictional forces and torques. Hence, usually work input > work output and $e < 100$ per cent.

Perpetual Motion

This limitation of the efficiency of a machine to no more than 100 per cent is very disappointing. A machine the energy output of which could be fed back to its input to make it run continuously, still leaving enough energy to spare to do some useful work on the side, would be a priceless gadget. It is no wonder that the invention of such a *perpetuum mobile* was the goal of numerous ingenious inventors in the past centuries. Figure 9.7 shows an example of such an engine. Heavy spheres S are attached to the rods R which are hinged at points H to a saw-toothed wheel. The idea of the inventor must have been about as follows: "The spheres on the right side of the wheel must always be farther from its center than the ones on the left. The torque which each of them exerts must be larger. This should start the wheel spinning clockwise. As it spins, the hinge mechanism will allow the spheres to move closer to its center as they pass from the right to the left side. Thus, the unbalanced torque and hence rotation must be maintained against frictional resistance and a possible external load."

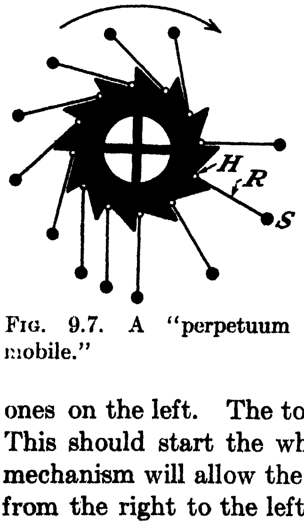


FIG. 9.7. A "perpetuum mobile."

Like all suggested perpetual motion machines, this one did not work. Why not?

Principle of the Impossibility of a "Perpetuum Mobile." The complete failure of such inventive efforts was finally accepted as the empirical basis of a physical principle: *The principle of the impossibility of a "perpetuum mobile."* **It is impossible to devise an engine which will deliver more work at one place than is put into it at another.** In other words, mechanical energy cannot be created out of nothing. Huygens writes on this subject in his "Horologium oscillatorium":

> If any number of weights be set in motion by the force of gravity, the common center of gravity of the weights as a whole cannot possibly rise higher than the place which it occupied when the motion began. . . . This hypothesis simply imparts what no one has ever denied, that heavy bodies do not move upwards—And truly if the devisers of new machines, who made such futile attempts to construct a perpetual motion would acquaint themselves with this principle, they could easily be brought to see their errors and to understand that the thing is utterly impossible by mechanical means.

(Note Huygens' caution in limiting his statement to mechanical means.)

Law of Conservation of Energy

The law of the impossibility of a perpetual motion (creation of mechanical energy out of nothing) is a special restricted version of the more general **law of conservation of energy** which states:

In transformations of mechanical or any other type of energy, energy can be neither created nor destroyed. Or to state it more impressively:

The energy of the universe remains constant.†

This seems at first glance to contradict our experience. When a sandbag drops from the fifth floor, its original potential energy is lost, and we don't see an equivalent ability to do work transferred to any other body. How can this observation be reconciled with the statement of the preceding law? This question will be answered in Chap. 10.

APPENDIX

Stability of Equilibrium

An important problem in connection with equilibrium is its stability; in other words, it is important to know whether or not the system tends to return to its state of established equilibrium after a small deviation from it.

For instance, the sphere in Fig. 9.8*a*, when displaced and released, tends to move back toward its original equilibrium position. This is the so-called **stable equilibrium.** In Fig. 9.8*b* it can be displaced by any amount from its original position and will remain at rest in its new position. This is the so-called **neutral equilibrium.** Finally,

---

† Dignified as this statement sounds, it may be perfectly meaningless. What do we have to know about the universe in order to decide whether or not this statement is meaningful?

Fig. 9.8*c* illustrates a case of **unstable equilibrium.** The ball is in equilibrium only as long as its center is precisely vertically above the apex of the hemispheric support. The slightest displacement will make it roll off. In other words, a small displacement will lead to a larger one, thus removing the object farther and farther from its equilibrium position.

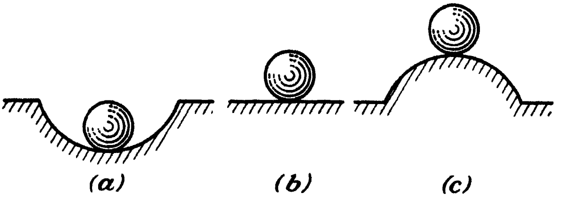


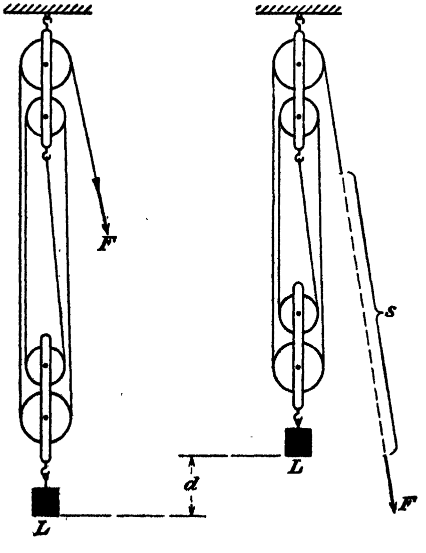
FIG. 9.8*a*. Stable equilibrium. *b*. Neutral equilibrium. *c*. Unstable equilibrium.

Generalization of the Concept of Stability

The concept of equilibrium is capable of generalization and will be encountered frequently beyond the field of mechanics. The terms of stability and instability are also used in a generalized sense. The meaning of equilibrium and stability in nonmechanical cases will become clear as we discuss particular problems. Figure 9.8*a* shows clearly a property which is common to all systems in stable equilibrium: Any attempt to shift the sphere from its equilibrium position raises its potential energy. We can generalize our definition of stability by stating: **A stable state is one of minimum potential energy.** This formulation makes it possible to extend the concept of stability to nonmechanical systems.

Example

Find the mechanical advantage of the system of pulleys shown in the figure, assuming it is to be frictionless.



Solution. According to the law of conservation of energy, the work done by the force F will be equal to the work done in lifting the load L:

$$Fs = Ld \tag{1}$$

where d = the height to which L is lifted and s = the distance through which the point of application of the force F is moved.

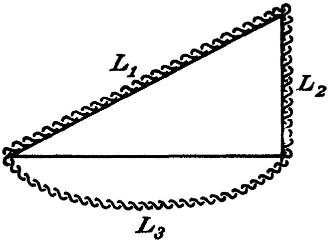
The mechanical advantage (MA) is defined as $MA = L/F$. Hence from equation (1) follows:

$$MA = \frac{s}{d} \tag{2}$$

To find the value of s for a given d, imagine the lower pulleys and the load L to be lifted through the distance d as indicated in the second diagram. By doing this, we create a slack in the ropes which originally supported the lower pulleys. Each of the four hanging ropes has to be shortened by *d* centimeters in order to remove the slack. This can be done by pulling the *F* end of the rope through the distance $s = 4d$. This determines the ratio $s/d = 4d/d = 4/1$. Hence $\underline{MA = 4}$.

QUESTIONS AND PROBLEMS

★**1.** The figure shows a typical proposed perpetual-motion machine consisting of a chain placed around a smooth inclined plane. Since L_1 is longer and, hence, heavier than L_2 while L_3 is symmetrical, the chain might be expected to keep revolving counterclockwise while doing work.



a. Restate the argument of the inventor.

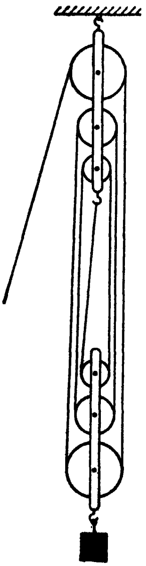
b. Show his fallacy.

c. How could you use the principle of the impossibility of a perpetual-motion machine to derive the equilibrium conditions of the inclined plane?

2. Which of the following statements is correct? The principle of the impossibility of a perpetual-motion machine implies that (*a*) you cannot construct a machine which will lift a 10-lb weight by the application of a 2-lb force; (*b*) you cannot build a machine (even if you could avoid friction) which will keep running forever; (*c*) you cannot design a machine which will yield 10 ft-lb of work if you apply 2 ft-lb of work to it.

3. When you roll a barrel up an inclined plane, you are using the inclined plane as a simple machine to lift a load. Explain why the inclined plane can be considered as a simple machine, and find its mechanical advantage, neglecting friction.

4. Find the mechanical advantage of the frictionless pulley system shown in the figure.



5. Why is it more strenuous to walk 8 miles than to ride the same horizontal distance on a bicycle? (Hint: If you hold a piece of chalk firmly against your hip and press the chalk against a blackboard as you walk along the board, you will obtain a wavy line.)

6. A force of 10^5 dynes acts on a 10-gm body through a distance of 20 cm as the body accelerates.

a. How much work is done?

b. What velocity does the body acquire?

c. How large is its kinetic energy?

d. How high could it rise by virtue of its kinetic energy?

7. From how tall a building would a car have to fall in order to acquire the speed of 80 mi/hr? How large will be the kinetic energy of a 3,000-lb car under these conditions?

8. A 20-gm ball rolls up an inclined plane (angle from the horizontal $\alpha = 30°$) until it stops after rolling 60 cm along the plane:

a. How large was the original kinetic energy of the ball?

b. How large is its final potential energy?

c. How large is the sum of potential plus kinetic energy of the ball as it rolls upward?

9. A 64-lb sled starting from rest moving on smooth ice is accelerated by a constant force of 32 lb over a horizontal distance of 2 ft. What final speed does it acquire?

a. Solve this problem using kinematic and dynamic equations.

b. Solve it by using energy considerations.

c. What advantage does the energy method have as compared with method *a*?

10. A boat is pulled by means of a rope which makes an angle of 30° with the direction in which the boat is moved. How much work is done in moving the boat 100 ft if the tension in the rope is 400 lb?

11. A 20-gm mass is moving with an acceleration of 200 cm/sec^2. How much kinetic energy is the body gaining per centimeter?

12. A 10-gm bullet is fired from a 2-kg rifle issuing with a velocity of 500 m/sec. Find the ratio (*a*) of the momentum of the bullet to the recoil momentum of the rifle, (*b*) of the kinetic energy of the bullet to the kinetic energy imparted to the rifle.

13. A simple pendulum consisting of a mass of 10 gm attached to a thin thread 1 m long swings making an angular excursion of 30° to either side of the vertical.

a. How large is the maximum variation in potential energy of the pendulum bob?

b. How large is the velocity of the bob in its lowest position?

c. How large is the sum of potential plus kinetic energy when the pendulum string forms an angle of 10° with the vertical?

14. In a hotel fire a man jumps from the sixth floor to the ground from a height of 60 ft. He is caught in a net which is bulged down 6 ft under his impact. Compute the average force exerted by the net upon the 160-lb man.

CHAPTER 10

WORK AND ENERGY: 2

a. From what kind of experiments do we infer the law of conservation of mass?

b. Does the law of conservation of momentum imply the conservation only of its magnitude or also of the direction?

c. Can you retrace the sequence of observations and generalizations which led to the law of conservation of energy?

d. Is the law of conservation of energy applicable only to processes on the earth or does it hold for the entire universe? How do we know?

e. What is the value of conservation principles from the standpoint of economy of thinking?

f. Where does their weakness lie?

g. Through what process of inference did we arrive at the general law of conservation of energy?

h. Does the law of conservation of energy apply to systems affected by friction?

i. What is the criterion distinguishing perfectly elastic from imperfectly elastic impact? Does the law of conservation of energy hold for the perfectly inelastic impact?

j. What factors determine the kinetic energy of a rotating body?

k. How do you measure work done in rotating a body?

l. What happens to the kinetic energy of a car when it stops?

m. Can the conservation of mechanical energy in frictionless systems be deduced from Newton's laws of motion?

n. Why do you get the same result when you solve a problem once using Newton's laws of motion and then, as an alternative, using the law of conservation of energy?

o. How would you revise the law of conservation of mass in the light of Einstein's law, $E = mc^2$?

Review of Conservation Principles

We have become acquainted so far with several principles which may be classified under the common name of "conservation principles." They are generalizations of experience with certain physical quantities which have the property of remaining invariant in the course of physical interactions.

1. The **law of conservation of mass** (discovered by chemists) states that the mass of a closed system† remains unchanged in any physical or chemical process which may take place in that system. (This law has been modified by Einstein as we shall see later in this chapter.)

2. The **law of conservation of momentum** states that the *vector sum* of the momenta in a closed system cannot be changed by interaction between the bodies composing the system; *i.e.*, the resultant momentum of the system remains conserved *in the absence of external forces.*

---

† Isolated group of bodies.

3. The **law of conservation of angular momentum** in a closed system asserts that the angular momentum of a system remains unchanged in the absence of external torques.

4. The **law of conservation of mechanical energy in frictionless systems** states that, when one form of mechanical energy is converted into another form, no loss or gain of energy takes place.

A precursor of the general law of conservation of energy, law 4, was induced from numerous, more restricted conservation laws which hold only approximately with actual bodies but which were assumed to apply exactly to ideal frictionless mechanical systems. Examples of the above-mentioned restricted conservation laws are, for instance, the law of conservation of work as applied to ideal simple machines or the conservation of kinetic energy in the process of elastic impact which was discovered by Huygens. Thus, for example, if two elastic balls collide, the sum of their kinetic energies after the impact is the same as before the impact. This furnishes incidentally an alternative criterion to distinguish perfectly elastic impact from inelastic collisions: *Whereas the law of conservation of momentum holds for all types of impact, the law of conservation of kinetic energy applies only to perfectly elastic impact.*

We shall complete our review of conservation laws by citing the most versatile and most important one:

5. The **law of conservation of energy** is the most general of all the conservation principles. It not only applies to all branches of physics but underlies all natural phenomena of the organic as well as inorganic world. *The strength of this principle* lies in its applicability to a wide range of phenomena and in the fact that it can be used in cases where the mechanism of the processes involved is unknown. *Its weakness* lies in the fact that it does not provide the means of establishing the mechanism of a process. It is essentially a bookkeeping principle in the household of nature.

Examples of Application of the Law of Conservation of Energy to Mechanical Systems. Before a discussion of the general principle of conservation of energy, we shall consider a few examples illustrating transformations of energy in mechanical processes.

Rotational Kinetic Energy. Can we determine the kinetic energy stored in a rotating body if we know its moment of inertia I and angular velocity ω?

Let us consider a wheel (Fig. 10.1) put in accelerated rotational motion by applying a force F to a string wound about its circumference. If the force moves the end of the string through the distance s, the work done is Fs. The torque applied to the wheel is $FR = T$, which makes

$$F = \frac{T}{R} \tag{10.1}$$

The relationship between the linear motion of the end of the string and the rotation of the wheel is given by

$$s = R\theta \tag{3.3}$$

For uniformly accelerated angular motion starting from rest ($\omega_0 = 0$) we derived

$$\omega = \alpha t \tag{3.6a}$$

and

$$\theta = \tfrac{1}{2}\alpha t^2 \tag{3.7a}$$

We can now substitute these data in our expression for work:

$$\boxed{Fs = \left(\frac{T}{R}\right)(R\theta) = T\theta} \tag{10.2}$$

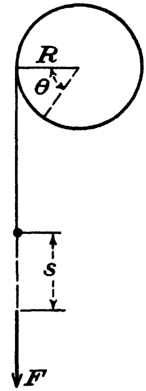


FIG. 10.1.

where $T/R = F$, $R\theta = s$. In words: **Work done in rotation is equal to torque times angle of rotation.**

Now, as a result of the applied constant torque, the rotation will be a uniformly accelerated one and θ will be given by (3.7*a*). The angular acceleration α in (3.7*a*) is determined by the law of rotational motion.

$$T = I\alpha \tag{8.8}$$

(where I is the moment of inertia of the wheel).

Hence, substituting (3.7*a*) and (8.8) in (10.2), we get

$$T\theta = (I\alpha)(\tfrac{1}{2}\alpha t^2) = \tfrac{1}{2}I(\alpha t)^2 \tag{10.3}$$

where $I\alpha = T$, $\frac{1}{2}\alpha t^2 = \theta$. Substituting (3.6*a*) in (10.3), we obtain

$$\boxed{T\theta = \tfrac{1}{2}I\omega^2 = \text{KE}} \tag{10.4}$$

Equation (10.4) is analogous to

$$Fs = \tfrac{1}{2}mv^2$$

Our equations tell us the following story: A torque T imparts an angular acceleration to a rotating wheel. When the torque is allowed to rotate the wheel through an angle θ (starting from rest), a final angular velocity ω is attained. This angular velocity determines the rotational kinetic energy of the body, KE $= \frac{1}{2}I\omega^2$. According to the law of conservation of mechanical energy, this kinetic energy is equal to $T\theta$, the rotational work done.

Rolling Motion. We are now in a position to use the law of conservation of mechanical energy to solve a problem which would have been much more complicated were we to solve it using Newton's laws of motion. The problem is as follows: A hoop rolls down an inclined plane, starting from rest. What final velocity does it acquire (Fig. 10.2)? Work is done on the hoop by the force of gravity:

$$\text{Work} = Wh = (mg)h$$

As a result the hoop acquires kinetic energy of translation (its center moves with a uniform final velocity v) and of rotation (it rolls with a constant angular velocity ω).

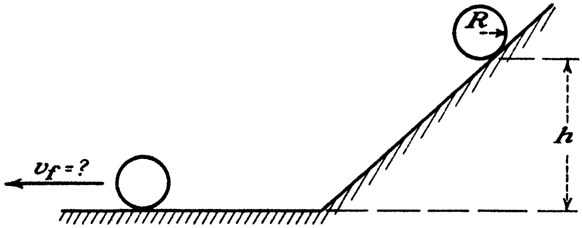


FIG. 10.2. Gain in kinetic energy of a descending hoop.

Our conservation law predicts the following relation:

$$\begin{aligned}\text{Work done} &= \text{kinetic energy acquired}\\ (mg)h &= \tfrac{1}{2}mv^2 + \tfrac{1}{2}I\omega^2\end{aligned} \tag{10.5}$$

where $\frac{1}{2}mv^2$ = translational KE, $\frac{1}{2}I\omega^2$ = rotational KE. *This single equation practically solves the problem.* We merely have to make appropriate substitutions using the relationships between I and m, v and ω.

For a thin hoop: $I = mR^2$ (8.7)
For rolling motion:[†] $s = R\theta$ (3.3)

Substituting (3.3) in the definition equation for speed, we get

$$\boxed{v = \frac{s}{t} = R\frac{\theta}{t} = R\omega} \tag{3.4}$$

We can now modify the term $\frac{1}{2}I\omega^2$ in equation (10.5) by setting $\omega = v/R$ from equation (3.4)[†] and $I = mR^2$ from equation (8.7). We thus obtain

---

[†] When a wheel rotates through an angle of 360° = 2π radians, it advances through the distance of $2\pi R$ centimeters. For any other angle $\theta = K(2\pi)$ radians, the distance covered will be $s = K(2\pi R) = (K2\pi)R = \theta R$ centimeters.

$$mgh = \frac{1}{2}mv^2 + \frac{1}{2}mR^2\left(\frac{v}{R}\right)^2 \tag{10.6}$$

$$gh = \tfrac{1}{2}(v^2 + v^2) = v^2$$

and finally

$$\boxed{v = \sqrt{gh}} \tag{10.6a}$$

It is interesting to compare this velocity with the one the hoop would have acquired by *sliding without rotation rather than rolling* down the smooth incline. In this case $\omega = 0$. Hence, equation (10.5) simplifies in this case to

$$mgh = \tfrac{1}{2}mv^2$$

and

$$v = \sqrt{2gh}$$

Hence, the final velocity in rolling is smaller than in sliding motion. (How does the moment of inertia affect the final velocity of a body rolling down an incline?)

Sliding Friction. Finally, to demonstrate the power of the conservation principle in reducing calculation, we shall solve a problem first by using Newton's laws and then by applying the energy method.

Problem. A block is pushed along a rough surface (Fig. 10.3). It is given an initial velocity $v_0 = 3.2$ ft/sec. Its weight is unknown, but we know the coefficient of friction $\mu = 0.32$.

How far does the block slide until it stops?

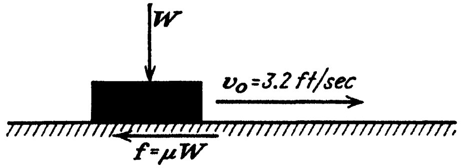


FIG. 10.3. Frictional retardation of a sliding block.

Solution. **First Method** (using Newton's laws of motion): The frictional force f which retards a moving body is given by the expression $f = \mu N$. N is the so-called *normal force*, which is the force with which the sliding body is pressed perpendicularly against the surface on which it slides. μ, **the coefficient of friction,**† is a constant which depends on the nature of the contacting surfaces. In the case of a body resting on a horizontal plane $N = W$, where W is the weight of the body, and we can write:

$$f = \mu W \tag{1}$$

---

† If f is the force which is just large enough to start the body sliding, μ is the coefficient of starting friction. If f is the force required to keep the body moving at constant speed, μ is the coefficient of sliding friction, which is somewhat smaller than the coefficient of starting friction.

In the situation pictured in Fig. 10.3, the motion is a uniformly decelerated one. The distance covered is given by

$$s = v_0t + \tfrac{1}{2}at^2 \qquad \text{(where } a < 0\text{)} \tag{2}$$

The relation between the deceleration and the frictional forces is given by Newton's second law:

$$f = ma = \frac{W}{g}a \tag{3}$$

Substituting (1) in (3), we get

$$f = \mu W = \frac{W}{g}a$$

Hence,

$$a = \mu g \tag{4}$$

For a decelerated body which eventually comes to rest ($v_f = 0$), Galileo's equation (2.5) can be written

$$v_f = 0 = v_0 + at \tag{5}$$

Substituting (4) in (5), we obtain

$$\underline{v_0} = -at = \underline{-\mu gt}$$

or

$$t = -\frac{v_0}{\mu g} \tag{6}$$

Substituting (4) and (6) in (2), we get

$$s = v_0\underset{\substack{\uparrow\\ t}}{\left(-\frac{v_0}{\mu g}\right)} + \frac{1}{2}\underset{\substack{\uparrow\\ a}}{(\mu g)}\underset{\substack{\uparrow\\ t^2}}{\left(\frac{v_0^2}{(\mu g)^2}\right)}$$

thus

$$\boxed{s = -\frac{v_0^2}{\mu g} + \frac{1}{2}\frac{v_0^2}{\mu g} = -\frac{v_0^2}{2\mu g}} \tag{7}$$

[The negative sign of s means merely that the direction of the displacement vector $\overrightarrow{s}$ (to the right in Fig. 10.3) is opposite to the acceleration vector ($\overleftarrow{a} = \mu g$ pointing to the left). Since we have chosen a plus sign for a in our equations, s is negative. Had we chosen a negative sign for the retardation a, s would have come out positive.]

The numerical value: $\underline{s} = -\dfrac{(3.2)^2}{2 \times (0.32) \times 32} = -\dfrac{1\,(3.2)^2}{2\,(3.2)^2} = \underline{-0.5 \text{ ft}}$

Second method (using the law of conservation of energy): Work done against the frictional force: Work $= fs$.
The kinetic energy dissipated:

$$\text{KE} = \frac{1}{2}mv_0^2 = \frac{W}{2g}v_0^2$$

According to the law of conservation of energy, the loss of kinetic energy is equal to the work done:

$$fs = \frac{W}{2g} v_0^2 \tag{1}$$

$$(\mu W)s = \frac{W}{2g} v_0^2 \tag{1a}$$

or

$$s = \frac{v_0^2}{2g\mu}$$

which is identical with the solution obtained by the first method. The first method required writing seven equations, the energy method only one, namely, (1*a*). The value of the conservation principle from the standpoint of economy of thinking is obvious, and the student would do well to investigate whether or not the energy method could be applied to a particular problem confronting him before proceeding to solve it by other methods.

Heat as a Form of Energy

Let us now return to the example of the falling sandbag of Chap. 9. What happens to its kinetic energy as it strikes the ground? The kinetic energy is gone, but something else has made its appearance. The ground and the sand have become slightly warmer. The disappearance of kinetic energy is accompanied by appearance of heat (this can be demonstrated more strikingly by hammering a nail repeatedly until it gets noticeably hot). Can we preserve the validity of the law of conservation of energy in this case by classifying heat as a form of energy and saying that as much energy was produced in the form of heat as was dissipated in the form of mechanical energy?

If we do so, we should interpret the above observation by saying: The kinetic energy of the falling hammer is not gone when it strikes the nail; it is merely transformed into an equivalent amount of a different form of energy, namely, heat.

To justify the extension of the law of conservation of energy by adding heat to the known forms of energy, we should have to make certain

1. That heat satisfies our criterion of a form of energy, *i.e.*, we should be able to transform heat into work.
2. That the amount of work produced from a given amount of heat is always the same and is equal to the amount of work which has to be expended to produce this amount of heat.

We shall call the number of ergs of mechanical energy produced from a unit† of heat energy the **mechanical equivalent of heat.** We shall have to confirm experimentally that, in all conversions of heat into mechanical energy and vice versa, the ratio of mechanical energy to heat is constant and is equal to the mechanical equivalent of heat.†

---

† The unit of heat energy, one calorie, is defined in Chap. 15.

The answer to our first criterion is well known to all. Ever since Hero of Alexandria, for about 2,000 years, steam engines of various kinds have been in use to transform heat into mechanical energy. Thus, heat is a form of energy by our definition. The law of conservation of energy was, however, not generalized to include heat for a long time in spite of the knowledge of basic facts on which such an extension could have been based. Investigations along the lines suggested by our second criterion occurred sporadically, and the crystallization of ideas proceeded gradually. Count Rumford, for instance, published in 1798 an account of his experiment on mechanical generation of heat. He was able to generate enough heat by drilling a steel plate with a dull drill to bring about 19 lb of water to the boiling point within 2 hr. He writes: " . . . and it appears to me to be extremely difficult, if not quite impossible, to form any distinct idea of anything capable of being excited and communicated in the manner in which heat was excited and communicated in these experiments, except it be motion. . . . " A clearer step toward the extension of the law of conservation of energy to heat is to be found in the writings of Sadi Carnot, the discoverer of the second law of thermodynamics. In the course of his reflections on the efficiency of steam engines, he conceived the idea of generalizing the principle of the impossibility of perpetual motion, extending it also to machines in which conversion of heat into mechanical energy took place, whereas previously its validity was known only for purely mechanical systems.

Carnot writes: "It will be objected here, perhaps, that a perpetual motion, proved impossible for purely mechanical actions is perhaps not so when the influence of heat or of electricity is employed. But can phenomena of heat or electricity be thought of as due to anything else than to certain motions of bodies, and as such must they not be subject to the general laws of mechanics?"

Mechanistic Philosophy in Physics

Carnot's statement is of twofold interest to us. (1) The impossibility of a perpetual-motion machine is synonymous with the impossibility of creating energy out of nothing. Carnot's argument implies that heat, just as mechanical energy, cannot be created out of nothing. (2) He attempts to justify his generalization by the assumption that "phenomena of heat and electricity are due to certain motion of bodies and hence must be subject to laws of mechanics."

---

† We rely for the time being on our intuitive conception of heat, leaving a more detailed discussion of the nature of heat, its conversion into mechanical energy, and its measurement to Chap. 16.

This second point is of particular interest, since it reflects the spirit of mechanistic philosophy in physics. This endeavor to explain all phenomena in physics by laws of mechanics is expressed even more strongly in the following quotation from Huygens' "Horologium": "In the true philosophy . . . the causes of all natural effects are conceived as mechanical causes. Which in my judgement must be accomplished or all hope of ever understanding physics renounced."

These significant statements forecast the program of theoretical physics of the nineteenth century. They set the aim of reducing all observed physical phenomena to mechanical explanations and of extending the laws of mechanics to all branches of physics. The law of conservation of energy is our first example of such an extension. Helmholtz showed that the validity of the law of conservation of energy could be derived for all branches of physics if we make the hypothesis that all matter consists of tiny mass points which interact by exerting "central forces"† upon each other and if we assume that all phenomena in all branches of physics can be explained by the action of such forces. We shall become acquainted subsequently with hypotheses modeled after the suggestion of Helmholtz which succeeded in making amenable to mechanical treatment various branches of physics, such as heat and electricity. In all such cases, conservation of energy is interpreted as conservation of mechanical energy of the interacting "mass points."

The branch of physics which deals with the relationship between heat and mechanics is called *thermodynamics.*‡ The law of conservation of energy is usually referred to as the **first law of thermodynamics.**

Further Generalization of the Law of Conservation of Energy

Once a general concept of energy was created, it was soon shown that heat and work were not the only forms of energy. For example, illumination of a black surface heats it; but heat can be converted into work; hence, light is a form of energy. Sound waves may cause material damage (*e.g.*, explosions) involving a noticeable release of energy. Electric motors offer a familiar example of conversion of electrical energy into work. Combustion of foodstuffs, which enables us to perform work, is an example of conversion of chemical into mechanical energy.

It is a curious fact that the law of conservation of energy was established primarily by three men who were not professional physicists. Two of them were medical men (Mayer and Helmholtz) and one a beer brewer (Joule). Mayer recognized heat, chemical energy, and light as forms of the same indestructible, transformable agent and computed the mechanical equivalent of heat from data on thermal behavior of gases (see Chap. 18). Helmholtz, independently of Mayer, rediscovered the principle somewhat later and worked it out in more detail and with somewhat greater generality. Joule made accurate determinations of the mechanical and electrical equivalents of heat.

---

† Forces between two particles whose direction coincides with the line joining the particles.

‡ Thermodynamics should be more appropriately defined as the study of interconversion of different forms of energy.

The latest accepted value for the *mechanical equivalent of heat* is **4.185 joules per calorie.**†

Is Mass a Form of Energy?

One of the greatest discoveries of this century was Einstein's deduction from his theory of relativity that mass and energy are two forms of the same thing and can be converted into each other. The law of conservation of mass thus has to be modified. Mass is not at all constant; it may be changed by certain physical processes (see Chap. 40), but the sum of mass plus energy remains constant.

The energy equivalent of mass is determined by **Einstein's law:**

$$\boxed{E = c^2m} \tag{10.7}$$

E is the amount of energy which appears when the mass m disappears or vice versa. *The proportionality constant* c^2, *or as we can call it, the* **energy equivalent of mass,** *is the square of the velocity of light.* (Its dimension is $(\text{cm/sec})^2 = (\text{erg/gm})$

Since $c = 3 \times 10^{10}$ cm/sec
we obtain from (10.7) $E_{\text{ergs}} = 9 \times 10^{20}_{(\text{cm/sec})^2} m_{\text{gm}}$

We see that even the conversion of a small mass should release an enormous amount of energy. Such transformation of mass is the source of energy in nuclear fission bombs as well as in stellar bodies such as the sun.

The use of Einstein's equation by nuclear physicists is now as common as the use of the mechanical equivalent of heat by power engineers.

Power

In the practical use of energy power is an important concept. *Power measures the rate of doing work* or, better, **the rate of transformation of energy.**

$$\boxed{P = \frac{E}{t}} \tag{10.8}$$

---

† For the definition of calorie see Chap. 16.

where P = the power, *i.e.*, energy transformed per unit time;
E = the amount of energy converted from one form into another;
t = the time required for the above conversion.

Equation (10.8) can be put in another practically useful form:

$$\boxed{P = \frac{E}{t} = \frac{Fs}{t} = F\frac{s}{t} = Fv} \tag{10.9}$$

This equation states that we find the power, *i.e.*, the rate at which work is done on a body, by multiplying the applied force by the velocity of the body.

We know from equation (10.2) that for rotational motion

$$\text{Work} = T\theta$$

From this we can deduce the equation for power in rotational motion:

$$\underline{\text{Power}} = \frac{T\theta}{t} = T\frac{\theta}{t} = \underline{T\omega} \tag{10.9a}$$

Action

We shall now add to our list of concepts another one which plays an important role in modern physics (see Chaps. 35 to 39). It is the concept of action.

Work: E = force × distance = Fs
Impulse: i = force × time = Ft
Action: A = force × distance × time = Fst

The concept of action is thus defined symmetrically with respect to time and distance, whereas work and impulse are space product and time product of force, respectively.

We can also reinterpret A as follows:

$$\begin{array}{ccccl} \underline{A = Fst} = & (Fs)t & = & (Ft)s & \qquad (10.10a) \\ & \uparrow & & \uparrow & \\ \underline{A =} & \underline{Et} & \underline{=} & \underline{is} & \qquad (10.10b) \\ & \uparrow & & \uparrow & \\ & \text{energy} & & \text{impulse} & \end{array}$$

Action is thus also energy times time or impulse times distance.

The product of force and distance occurs also in the definition of a force moment (torque), and thus action is seen also to have the dimension of torque times time:

$$A = (Fs)t = Tt = \text{torque} \times \text{time} \tag{10.10c}$$

and since

$$Tt = \Delta I\omega \tag{8.9}$$

we see that action has the same physical dimension (see Chap. 6) as the angular momentum. This fact will be considered in the derivation of the equations of Bohr's atomic theory in Chap. 36.

APPENDIX

Units of Energy and Power

Units of Energy

1 erg = 1 dyne-cm
1 joule = 10^7 ergs = 0.239 cal
1 ft-lb = 1.36 joules
1 cal = 4.19 joules = 3.09 ft-lb

Units of Power

1 watt = 1 joule/sec = 0.239 cal/sec
1 hp = 550 ft-lb/sec = 745 watts

QUESTIONS AND PROBLEMS

1. A 20-gm stone is shot from a slingshot straight upward. The average force exerted upon the stone is 10^6 dynes, and it acts upon the stone through the distance of 10 cm. How high does the stone rise?

2. A stone is shot from the slingshot of Prob. 1 in a direction of 60° up from the horizontal. It describes a parabola. How high is the apex of this parabolic path above the slingshot?

3. A hoop having a mass of 500 gm and a radius of 50 cm is rolling at a speed of 40 cm/sec on a horizontal plane. How high will it roll up an inclined plane of 30° slope?

4. Two identical spheres are released from the same height of an inclined plane simultaneously and move downward until they reach the horizontal plane at the bottom; assume the sphere *A* to slide without rolling (no friction) and the sphere *B* to roll normally (without slipping). Which of the following statements is correct?

a. Sphere *A* arrives first at the bottom because no rotational energy is stored in it.

b. Sphere *B* possesses a larger total kinetic energy than *A* when both arrive at the bottom because of its rotation.

c. Both spheres have the same kinetic energy of linear translation when they arrive at the bottom.

d. None of the preceding statements is correct.

5. A 16-lb block slides up 5 ft along an inclined plane, rising 3 ft in the process. Its initial velocity is 20 ft/sec.

a. What is its initial kinetic energy?

b. What is the minimum loss in kinetic energy it will suffer?

c. What is the maximum possible value for its final kinetic energy?

d. If you actually find its final velocity to be 10 ft/sec, how would you explain it?

6. It requires an expenditure of 40 ft-lb of energy per foot of displacement to move a box weighing 320 lb across a horizontal floor with constant speed. How large is the coefficient of friction between the box and the floor?

***7.** Assume the force in Prob. 6 to be applied as a pull transmitted through a rope inclined at an angle of 30° toward the horizontal (the prolongation of the rope passing through the center of gravity of the box). How large must this force be to impart to the box the velocity of 2 ft/sec after the box has moved 6 ft?

***8.** Solve Prob. 7 assuming the box to be pushed by a rod rather than pulled by a rope.

9. A 20-lb box slides 10 ft along a horizontal table until it stops. The coefficient of friction between the box and the table is 0.2.

a. Find the original speed of the box.

b. Find the retarding force.

c. How many calories of heat are generated in this process?

10. Does high instantaneous power imply generation of a large amount of energy?

11. A 100-lb eagle is seen to rise with a vertical velocity of 20 ft/sec. How large a horsepower does it develop in this flight?

12. A 60-lb box moves with constant speed along a horizontal table. Assuming the speed to be 2 ft/sec and the coefficient of friction to be 0.5, compute the power necessary to keep the box moving at this constant speed (*a*) in foot-pounds per second, (*b*) in horsepower.

13. A 140-lb cyclist riding a 20-lb bicycle requires 2 min in order to reach the top of a mountain road of uniform slope starting from the bottom. The road is 400 ft long, and the slope is 30° against the horizontal. How much power does he develop? (Neglect friction.)

14. An organ-grinder turns the crank of his organ at the rate of $\frac{1}{2}$ revolution per second, applying to the crank handle a constant force of 0.02 lb at right angles to the radius, which is 1 ft long.

a. How much power does the operation of the organ require?

b. How much work does he do in 1 hr?

15. On the basis of what evidence do we classify heat as a form of energy?

16. Give examples illustrating transformations between different forms of energy.

CHAPTER 11

FLUIDS AT REST

a. What properties do liquids and gases have in common?

b. In what way do liquids differ from gases?

c. Is water compressible?

d. What does "to flow" mean?

e. What is pressure?

f. Why is a scalpel sharp?

g. Why do human tissues feel soft?

h. How could you determine the hydrostatic force exerted on a surface if the hydrostatic pressure is given?

i. What is the direction of hydrostatic force against a solid wall?

j. What does Pascal's principle state?

k. What physical principle do we use to derive Pascal's principle?

l. Does the blood pressure in the feet of a standing man change when he lies down?

m. How does the hydrostatic pressure at a certain depth depend on the shape of the container?

n. What is the difference between the buoyant force exerted in water (1) upon an air bubble and (2) upon a lead sphere of equal volume?

o. Discuss the principles used in our derivation of Archimedes' principle.

p. What criterion do you use to decide whether or not a body will float?

q. Will a block of aluminum sink in water as deeply as an equally large block of lead?

r. (1) Why are falling drops almost spherical? (2) Why are they not perfectly spherical? (3) Under what conditions will a falling drop be perfectly spherical?

s. How is it possible for some insects to run along the surface of water as easily as over solid ground?

Classification of States of Matter

We classify matter according to its **"phase"** as follows:

1. **Solids** have size and shape.
2. **Fluids**
    - *a.* **Liquids** have size but no definite shape.
    - *b.* **Gases** have neither a definite size nor shape.

1. **Solids** can be deformed by forces, and their size may, for instance, shrink under pressure. They have a tendency to resume their original size and shape when the deforming forces are removed.

2. **Fluids** are materials which can flow. Flow is a motion through constraints accompanied by deformation. For instance, water flowing out of the bottle of Fig. 11.1 is deformed: Moving through the part *A* water moves in the shape of a wide cylinder, while to pass through part *B*

it is deformed into a narrow one. Solids, however, can also be made to flow when they are powdered.

2a. A partly confined liquid assumes the shape of the container, and its free surface assumes the shape of an equipotential surface. If this were not the shape of the surface, a liquid particle would tend to move from points of higher to those of lower potential energy until stability is reached, when all neighbor particles have equal potential energy.

2b. While liquids assume the shape but do not necessarily fill the size of the container, gases fill any confined space offered to them.

Compressibility of Liquids

Gases are highly compressible. Liquids, on the other hand, have been for a long time regarded as incompressible. Experiments have been performed in Florence which seemed to prove it. Water was sealed in a spherical iron bomb. Any deformation of such an iron bomb by hammering should have reduced the volume of the container and, hence, compressed the water. One actually succeeded in reducing the volume of the container, but water was "sweated out" through pores in the texture of the metal. Thus, the attempt to compress water proved a failure. Later on, however, Oersted, using a refined technique, succeeded in demonstrating the fact that liquids are slightly compressible. The smallness of this compressibility is illustrated by the fact that a volume of water originally under atmospheric pressure suffers a volume shrinkage of only 1/22,000 of its original volume when the pressure is doubled.

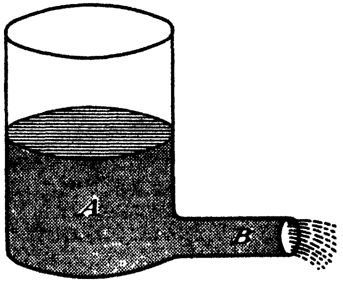


FIG. 11.1. Outflow of liquid from a tank.

Definition of Pressure. Let us now find an appropriate definition for the concept of pressure which we just used, relying on our intuitive (rather loose) conception of it. Pressure is measuring, so to speak, the "concentration" of the acting forces. We might have also called it "force density." It is measured by the quotient

$$P = \frac{F}{A} \tag{11.1}$$

(where F is the resultant of the acting forces perpendicular to the area A).

In words: **Pressure is equal to force per unit area.**

This means that the average pressure on any plane surface is equal to the total force on the surface in a direction normal to the surface divided by the area of the surface.

The effect of forces upon matter is determined not only by their magnitude but also by their concentration, *i.e.*, by the pressure. Why is the surgeon's scalpel sharper than a butter knife? Because its cutting edge has a smaller area of cross section. This enables the surgeon by exertion of a moderate force to subject the tissue to be cut to a very high local pressure. For the same reason a sharp injection needle can penetrate tissues more easily than a dull one. On the other hand, there are situations in which it is desirable to avoid high pressures. For instance, the purpose of snowshoes is to distribute the pressure due to body weight over a large area in order to avoid sinking into the snow.

Why is a feather bed more comfortable than a cement floor? Because only a few points of your body come in contact with the hard floor and a high pressure at a few points causes discomfort. On the other hand, a soft bed is one which makes contact with one's body over a large area, thus minimizing the load per unit area.

The concept of pressure plays an important role in the study of fluids. The pressure in a vessel containing liquid (**hydrostatic pressure**) increases with depth as can easily be shown. For instance, the size of a gas-filled rubber balloon or of gas bubbles can be seen to increase as they rise from great depth to the water surface.

Direction of Hydrostatic Forces

The force due to hydrostatic pressure exerted on a solid wall by a liquid *at rest* is always perpendicular to the wall. Why? Imagine a small volume element e (a blob of liquid) in a liquid *at rest*. Assume, whether it be right or wrong, that the pressure forces which it exerts on the wall are oblique and give rise to a resultant force F (Fig. 11.2). According to Newton's third law, the wall should exert an opposite reaction R upon the fluid element. This reaction could be resolved into two components: one perpendicular to the wall (R_x) and one parallel to it (R_y). The parallel component R_y will accelerate our blob of liquid parallel to the wall. But then our fluid will not be at rest, contrary to our assumption! Hence, we must conclude that, in a fluid at rest, R_y must be zero; in other words, R must be perpendicular to the wall.

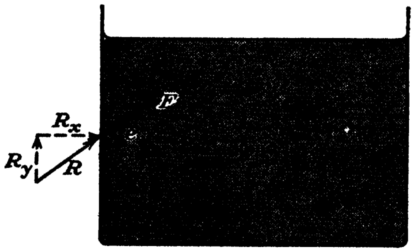


FIG. 11.2.

Hydrostatic Pressure as a Scalar

It means no more to speak of the direction of hydrostatic pressure at a point than to speak of the direction of a temperature. This can be shown as follows: Fig. 11.3 shows a cylinder with a piston which compresses the enclosed fluid. For simplicity, we imagine the system removed from the influence of gravity. (The reason for this restriction will become clear later.) Suppose you decide to define pressure p as a vector equal to force divided by an area perpendicular to the force. Then the p vectors exerted upon the surface of the small spherical volume element e of the fluid will converge toward the sphere's center. Suppose we now allow the sphere to become infinitely small, *i.e.*, we allow it to shrink into a point; what will be then the direction of the pressure (defined as above) at that point? It should point at once in all directions. Thus, obviously the above concept of a hydrostatic vector pressure is not tenable, and therefore, **hydrostatic pressure is defined as a scalar.** (Since F is a vector, the equation $F = pA$ implies that the area A should be defined as a vector, so that $\overrightarrow{F} = p\overrightarrow{A}$.)

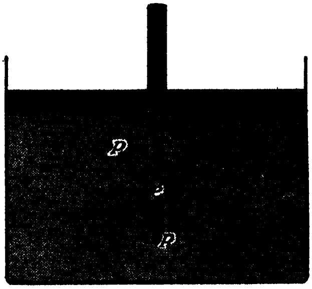


Fig. 11.3. Spherical surface subjected to hydrostatic pressure.

Pascal's Principle

Let us use our fluid which is confined in a *closed* cylinder in a region free of gravity to derive **Pascal's principle,** which states that **if you change the pressure at one point in a confined fluid at rest, you will produce the same pressure change at all other points.**†

To deduce this statement, let us consider an imaginary cylindrical fluid element AB in a fluid at rest (Fig. 11.4). The hydrostatic-pressure forces H exerted upon its curved surface must cancel out for symmetry reasons. The force on the left plane surface A is

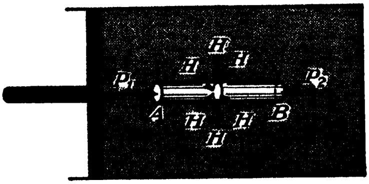


Fig. 11.4.

$$F_1 = P_1a$$

and on the right surface

$$F_2 = P_2a$$

(a is the cross-sectional area of the cylinder. P_1 and P_2 are the values of the hydrostatic pressure at A and B, respectively.) From this, R the resultant force acting on our cylinder, can be found as follows: $R = F_1 - F_2 = P_1a - P_2a = a(P_1 - P_2)$. But this resultant would accelerate our cylinder, which contradicts our assumption of a fluid at rest. Hence, we have to assume that, in a fluid at rest,

---

† Pascal's principle does not apply to pressure changes which are due to changes in weight per unit volume or due to acceleration of the fluid.

$P_1 = P_2$. If we now change P_1 by ΔP, P_2 will have to change by the same amount. In other words, a pressure change at one point will cause the pressure to change by the same amount everywhere else.

The hydraulic press (Fig. 11.5), the principle of which finds wide use in devices such as the hydraulic brake, the hydraulic jack, the hydraulic lift of operating room tables and dentist's chairs, etc., illustrates the meaning of Pascal's principle. The apparatus shown in Fig. 11.5 is essentially a U tube with a wide leg L (of cross-sectional area A) and a leg R of smaller cross-sectional area a. A force f applied to the small plunger will give rise to a pressure increment $p = f/a$. This pressure will be propagated throughout the liquid and will give rise to an upward force F exerted upon the large plunger of area A: $F = pA = (f/a)A$. From this equation, we can infer that the following proportion should hold between the force F (which is equal to the load W held up by the press in equilibrium), the applied force f, and the cross-sectional area of the cylinders R and L:

$$\frac{F}{f} = \frac{A}{a}$$

Since we are at liberty to make the ratio A/a of the areas as large as we please, this machine can provide a great mechanical advantage, allowing a small force f to lift a weight $W = F = (A/a)f$.

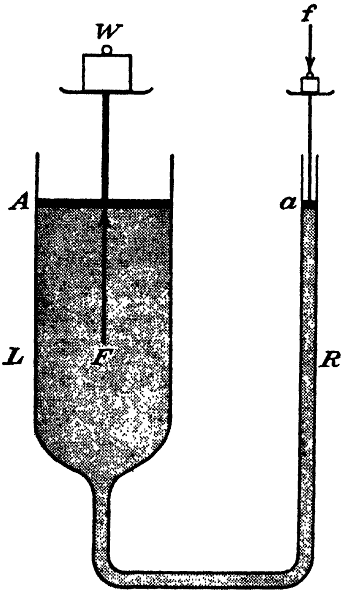


FIG. 11.5. Hydraulic press.

Magnitude of the Hydrostatic Pressure

Now let us consider the effect of gravity on a liquid at rest. We have mentioned before that the pressure in a liquid increases with depth. This is due to gravity. Imagine a small horizontal area A at the depth h in the liquid (Fig. 11.6a).

The pressure exerted on it is equal to the weight of the overlying liquid divided by the area.

$$\boxed{p = \frac{W}{A} = \frac{Vdg}{A} = \frac{(Ah)dg}{A} = hdg} \tag{11.2}$$

[V = the cylindrical volume equal to (Ah), d = the density equal to mass per unit volume, dg = the weight per unit volume.]

We see the following from (11.2): (1) p depends on the depth h and is the same for all points in the same horizontal plane (h = constant), (2) p depends on the density d of the liquid.†

Does this result hold only for a cylindrical container or for any shape of the vessel? Is the pressure at P in Fig. 11.6*b* the same as in Fig. 11.6*a*? This can be answered by making use of the assumption that the pressure at a point in a fluid is the same whether the boundary passing through that point is a fluid-fluid or fluid-solid interface. The justification of this assumption is provided by the correctness of the predictions which are based on it. In Fig. 11.7*a*, the pressure in the tank at the point P is equal to hdg. Now draw an *imaginary* boundary wall WW, conceiving for clarity the liquid on the right side of the wall to be colored (shaded area). According to our assumption, we can imagine the boundary between the colorless and the colored parts of the liquid to be frozen solid without affecting the pressure at any point along the boundary. We can now remove the colored liquid, thus obtaining a container with an irregular solid wall WW (Fig. 11.7*b*).

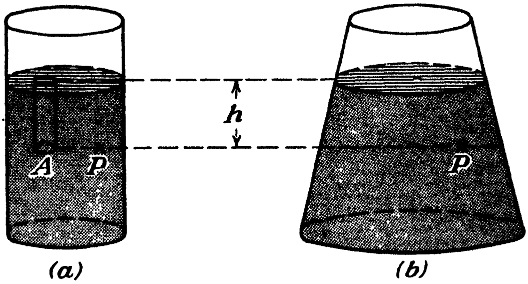


FIG. 11.6.

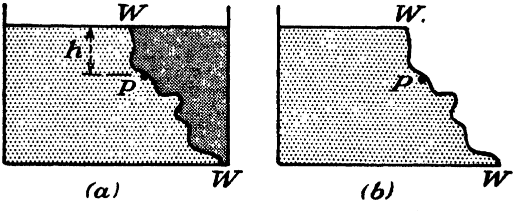


FIG. 11.7*a*. Illustration for determination of hydrostatic pressure at point P in the interior of a liquid. *b*. Illustration for determination of hydrostatic pressure at point P at an irregular wall.

---

† The pressure of the atmosphere is explained similarly. In the case of the atmosphere, however, there is no sharp upper surface from which the depth could be measured and the density d is not constant but, owing to the great compressibility of the gas, increases with depth.

The pressure at any point of this wall will be the same as it was before the "freezing" process, namely, $p = hdg$. Hence, since we are at liberty to choose any shape of the wall W, and since we can pass an imaginary boundary through any point, we see that the pressure depends only on depth and not at all on the shape of the container. Thus, for instance, the pressure at point P is the same in Figs. 11.6*a* and *b*.

The method of "freezing" imaginary boundaries is very useful in deriving hydrostatic theorems. We want to use it now to show that the liquid levels stand at an equal height in a U tube, regardless of the width of its arms.

Consider a tank filled with a liquid (Fig. 11.8). In this tank visualize an imaginary boundary (dotted lines) representing a U tube with arms of different width. You can imagine that the liquid on the outside of the U tube acts as a container wall. The function of this liquid wall is to provide the wall pressure required to keep the U-shaped inner fluid in equilibrium. A solid wall could do this just as well as a liquid wall; you can thus imagine the dotted boundary frozen solid (and, if you wish, all the liquid on the outside as well), without disturbing the equilibrium of the inner liquid. You have obtained in this fashion a solid U tube, and you see that in equilibrium the liquid level is the same in both of its arms despite unequal width.

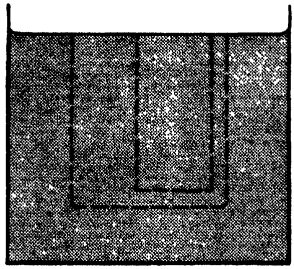


FIG. 11.8. Equilibrium of liquid in a U tube.

Measurement of Blood Pressure

The understanding of liquid equilibrium in U tubes is of importance for physicians and physiologists, since U-tube manometers are used for measurement of blood pressure. In the simplest case, such a manometer consists of a U tube of uniform diameter filled with a liquid, *e.g.*, mercury. As long as the two open legs are subject to atmospheric pressure, the liquid level is equal in both manometer arms. If, however, one of the two arms is connected to a reservoir whose pressure exceeds the atmospheric pressure (for instance, a blood vessel), the liquid level will be depressed in the "high-pressure arm" and will rise in the open arm. The difference between the two liquid levels (H) multiplied by the weight per unit volume (dg) of the liquid measures the excess of the pressure in the reservoir over the atmospheric pressure.

As an example we shall consider measurement of blood pressure in an animal with the aid of a U-tube manometer M (Fig. 11.9).† There are two complicating features which must be borne in mind in this case: (1) The pressure must be transmitted from the animal to the manometer through a sodium citrate solution (to avoid clotting of the blood), and (2) the manometer and animal may be on two different levels, which must be taken into account in computing the blood pressure. It would be quite wrong to take the manometer reading H as an indication of the blood pressure at P. The correct value of p follows from the following consideration. Draw a horizontal line AB, and consider the mercury below it to act as a kind of a scale. The pressure on the left "pan" at A is $p_L = HDg + p_A$ (where D is the density of the mercury and p_A the atmospheric pressure). The pressure on the right "pan" is also due to two factors: (1) hydrostatic pressure of sodium citrate solution (density d) which equals hdg, and (2) blood pressure p which is transmitted according to Pascal's principle through the citrate solution and acts on "pan" B.

---

† Such setups are commonly used for measurements of arterial pressure. In this example we shall disregard the fact that the pressure is pulsating.

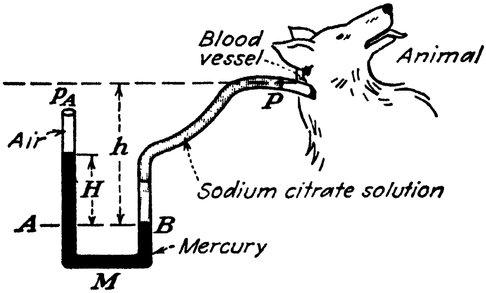


FIG. 11.9. Setup for measurement of blood pressure in an animal.

In equilibrium:

$$p_A + HDg = hdg + p$$

or

$$\boxed{p - p_A = HDg - hdg} = \Delta p$$

We see that the reading H of the mercury column has to be "corrected" by a term depending on the density d of the citrate solution and on the relative position of animal and manometer (more precisely, on the height h of the blood vessel above the manometer level B).[†]

In case the pressure is transmitted through air instead of the sodium citrate solution, we can practically set $d = 0$ and the correction becomes unnecessary. Such is the case in clinical measurements of human blood pressure where a manometer measures the air pressure in a "compressed-air cuff." In this case the simplified forumla $\Delta p = HDg$ is used. Evidently, Δp determined in Fig. 11.9 is not the actual blood pressure but its excess over the atmospheric pressure.†

[†] Water rather than mercury is used as a manometric fluid for measurements of venous pressure which is rather low. For a given pressure the water manometer gives a much larger H than a mercury manometer.

Archimedes' Principle and Buoyancy

Our discussion of equilibrium of liquid particles must have suggested to the student a disturbing idea. Consider a blob of liquid (this consideration applies to gases as well as to liquids) in a container (Fig. 11.10). In a fluid at rest all particles are motionless by definition, and yet we should expect our blob of liquid to fall, since it has a weight W! What prevents it from falling? Besides gravity there is only one other source of force action upon our fluid particle: the hydrostatic pressure exerted upon it by its environment (see the hydrostatic-pressure force vectors in Fig. 11.10). These hydrostatic-pressure forces ($F = pA$), however, are not the same at all points of the surface of our blob. The pressure increases with depth! As the diagram suggests, the resultant of the pressure forces cannot be expected to vanish. We must rather expect a resultant upward force B. *The fact that the particle remains at rest shows that this upward resultant is equal to the weight of the particle* ($B = W$.) We shall call this resultant force **buoyancy. The force of buoyancy is thus equal to the weight of the displaced fluid.** (This statement is justified since our blob displaces its own volume of fluid.)

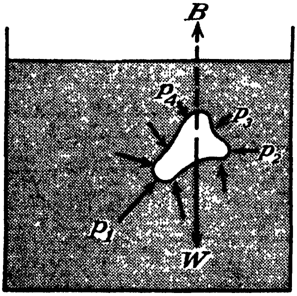


FIG. 11.10. Illustration for the derivation of Archimedes' principle.

Now we imagine the fictitious boundary of our fluid blob frozen solid. We then withdraw the fluid from the interior of this imaginary shell. The only force acting on the shell now is buoyancy. We thus see that an empty "vacuum bubble" is acted upon by an upward buoyant force which is equal to the weight of the displaced fluid.

If we decide to fill the emptied space of volume V with a different substance, say lead, the lead particle will be under the action of two forces: (1) its own weight $\downarrow W = m_L g = VDg$ (m_L, mass of the lead; D, the density of the lead) and (2) buoyancy $\uparrow B = m_F g = Vdg$ (m_F, mass of the fluid displaced by the lead particle; d, density of the displaced fluid).

Thus, the resultant force is

$$\boxed{R = \downarrow W - \uparrow B = Vg(D - d)} \qquad (11.3)$$

---

† The internal pressure of animals and fish is adjusted to the pressure of their environment. Deep-sea fish, for instance, which live under enormous hydrostatic pressures, explode when caught and brought to the surface of the ocean.

where R is the apparent weight of the body when submerged in the fluid; *i.e.*, it is the force it would exert on a spring scale when fully submerged.

Equation 11.3 expresses the famous **principle of Archimedes,** which states

A submerged body is buoyed up by a force equal to the weight of the displaced fluid.

This apparent loss of weight is familiar to all. It accounts for the ease with which you can lift heavy stones under water and perform submarine gymnastic feats which you might not be able to do in air. This fact is utilized in medicine by prescribing underwater gymnastics to patients recovering from paralysis who are too weak to lift their limbs without the benefit of buoyancy.

It is important to note that the "apparent weight" R [equation (11.3)] may become zero and even negative. The former happens when $D = d$, which means that water has zero apparent weight under water; the latter occurs when $d > D$, in other words when the submerged body is less dense (at least on the average) than the fluid. A negative apparent weight is what we might call a "lift." It makes CO_2 bubbles rise in ginger ale and a zeppelin rise in the air. It is well to keep in mind that the lift of a zeppelin is due to the variation of the atmospheric pressure with depth. Without such a pressure variation there can be no buoyancy.

Flotation. When does a body float? What determines its depth of submergence? As a body sinks deeper and deeper under a liquid surface (Fig. 11.11), its weight remains constant but the buoyant force increases with increasing displacement of the liquid. Finally, if the body is less dense (on the average) than water, the buoyant force will become equal to the weight before the entire volume has been submerged. We have then the following simple equilibrium equation:

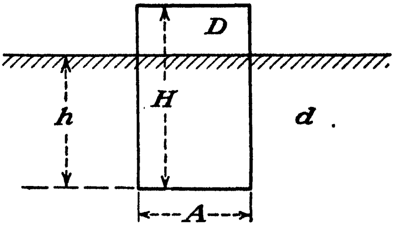


FIG. 11.11. Flotation of a prismatic body.

$$W = (AH)Dg = (Ah)dg = B \qquad (11.4)$$
$$\text{(Weight = buoyancy)}$$

as our condition of flotation (A = cross-sectional area; see Fig. 11.11). From equation (11.4) follows for a prismatic object of height H submerged in a liquid of density d (see Fig. 11.11):

$$\frac{H}{h} = \frac{d}{D}$$

This equation shows that the depth of submersion depends also on the density d of the liquid and suggests that comparative measurements of density may be made by observation of the depth of submergence. This principle is utilized in hydrometers, which are extensively used in biological laboratories for measurement of the specific gravity of solutions. A hydrometer is essentially a floating vertical tube provided with a scale on which the depth of submersion is read. The scale can be calibrated to read directly the salt concentration, the specific gravity of the liquid, etc.

Centripetal Buoyancy. Our equation for the apparent weight of a submerged body

$$R = V(D - d)g \tag{11.3}$$

shows that the body becomes heavier as g increases in cases where $D > d$ and more buoyant when $d > D$. Thus, if you want to precipitate particles whose density is very close to that of the liquid, $(D - d)g$ will be very small but could be increased by increasing the acceleration g. This can be accomplished by rapid rotation (**centrifuging**), which permits one to substitute for the gravitational acceleration g a much larger centripetal acceleration. Thus red blood cells, for instance, being heavier than plasma, can be made to drift with increased terminal velocity toward the bottom of the rotating test tube in a centrifuge (Fig. 6.5). For the fat globules in the milk, on the other hand, $d > D$. Rapid rotation may increase their buoyancy and make them move toward the axis of rotation more rapidly than they were moving toward the milk surface in a stationary container (cream separator).

APPENDIX I

Some Biological Applications of Archimedes' Principle

1. A Microrespirometer. This is an interesting example of a toy which has been turned into an important scientific tool of extreme sensitivity (Lindestrøm-Lang, 1937; Boell, Needham, and Rogers, 1939). In fact, the idea of this instrument occurred to Lindestrøm-Lang when he was playing with his children's toys. The toy which captured his imagination was the "cartesian diver" invented by the great philosopher René Descartes. Figure 11.12 illustrates the principle of this toy. A cylinder is filled with water and closed off by an elastic membrane M. Inside the cylinder is placed the diver, D, represented in our case by an inverted test tube partly filled with water and partly filled with air which forms a bubble at the top of the tube. The size of the bubble is such that the diver tends to float and touches the membrane. If, however, a sufficient pressure is exerted by hand upon the membrane, the diver can be made to sink to the bottom or to remain floating at a definite height in the cylinder.

How is the behavior of the diver to be explained? The inverted test tube "rides" on the air bubble; it is pushed upward by a buoyant force which is equal to the weight of the water displaced by the bubble. If the bubble is large, the buoyancy may exceed the "reduced weight" of the glass tube (*i.e.*, its vacuum weight minus the weight of the water the glass displaces) and the diver will rise. If, on the other hand, the bubble is small, its buoyancy may be smaller than the reduced weight of the glass and the diver will sink. There is a critical size of the bubble such that the diver will remain in equilibrium at a given height.

Now the size of the compressible bubble (whose mass is negligible against the mass of the diver) can be controlled by varying the pressure exerted by hand upon the membrane, which is propagated throughout the liquid according to Pascal's principle, and thus the diver can be made to rise, to sink, or to float.

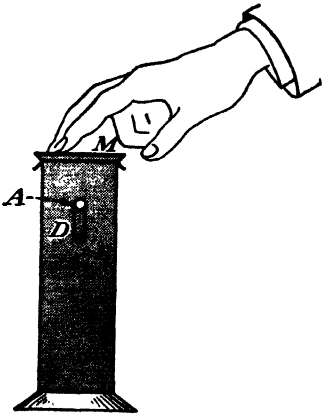
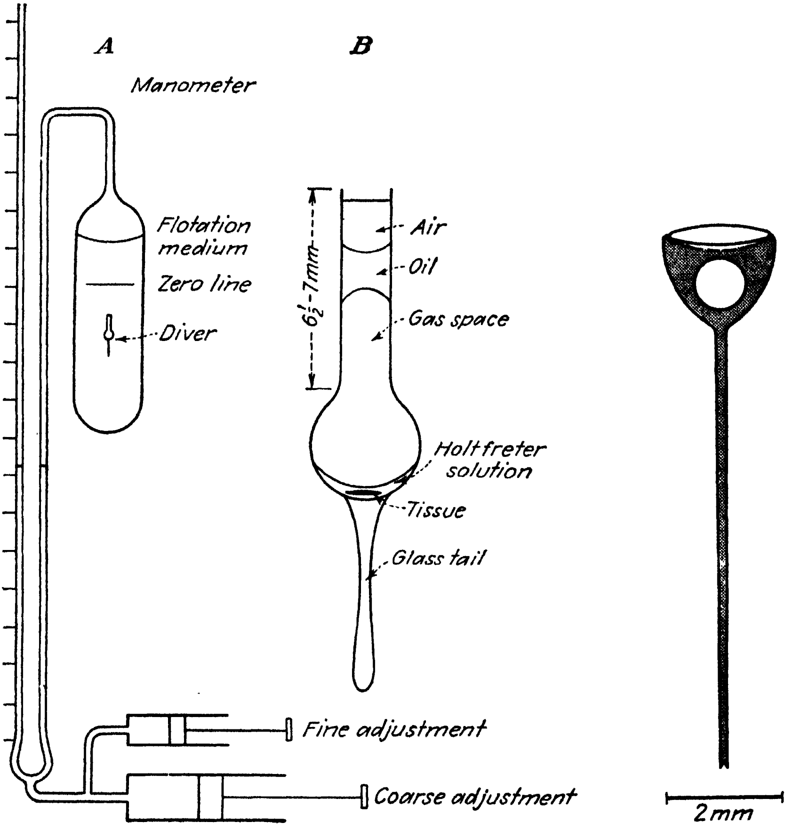


Fig. 11.12. Cartesian diver (*A*, air bubble).

The cartesian-diver principle opened a wide field of biological research, making it possible to study respiration of single cells and fragments of embryonic tissues. Figure 11.13*a* shows a cartesian-diver microrespirometer. The diver proper is shown in Part *B*. The tissue is placed in a physiological salt solution which is in contact with an air space which is closed off by an oil droplet. The purpose of the glass tail is to keep the diver in a vertical position. Part *A* shows the diver in a container which is filled with an appropriate liquid. This container is connected to a system of syringes and a manometer. By means of the syringe plunger, the height of the liquid in the manometer leg can be varied, and thus the pressure in the vessel containing the diver can be varied and measured. We now have essentially the same conditions as in Fig. 11.12. The diver is adjusted to begin with so that a mark on the diver coincides with a zero line marked on the large container. In the course of respiration, the tissue gives off CO_2, which is highly soluble in water and thus does not noticeably affect the gas volume in the diver. At the same time oxygen is consumed, which leads to a diminution of the air "bubble." As a result, the diver tends to sink, but it can be restored to its original level by lowering the pressure in the large container. Thus, by keeping the diver at a constant height, we keep the air bubble at constant volume. To keep the volume constant, we lower the pressure and read it on the manometer. From these pressure readings the rate of oxygen consumption can be easily determined (see gas laws, Chap. 18). This instrument is so sensitive that changes in gas volume in the order of 10^{-6} cm^3 (at standard conditions) can be measured. It was thus possible to study, for instance, the respiration of an isolated egg cell during the process of cell division (Zeuthen).

2. **A Microbalance.** Another ingenious use of the cartesian-diver principle was made by Zeuthen, who devised a balance sensitive enough to weigh a single ameba. In a study extending over several weeks, he recorded the steady diminution of weight of an ameba while subjecting the unfortunate animal to starvation.

Figure 11.13*b* shows a diagram of such a microbalance (from Zeuthen's paper). It is a cartesian diver with a hollow tail. The white circle indicates a spherical bubble. The ameba to be weighed is placed on the top of the diver, which serves as a scale pan. The adjustment of the equilibrium of the diver is accomplished by means similar to those used with the microrespirometer. It is to be kept in mind that what one measures with this microbalance is the "reduced weight" of the ameba, *i.e.*, its vacuum weight diminished by the force of buoyancy. Very tiny pieces of gold foil are used as standard weights.



(*a*) (*b*)

FIG. 11.13*a*. Cartesian-diver microrespirometer. (*From DeRobertis, Nowinski, and Saez, "General Cytology," Saunders.*) *b*. Cartesian-diver microbalance. (*Courtesy E. Zeuthen.*)

3. **Cushioning of Centrifuged Cells.** When reading about Harvey's experiments in Chap. 6 in which he subjected cells in a centrifuge to forces thousands of times as large as their normal weight, you may have wondered why the cells are not flattened and squashed by their own apparent weight. In fact, this is what would have happened if the cells were allowed to be centrifuged against the bottom of the test tube. But they are prevented from settling by the use of a sugar solution of suitable concentration which is poured into the sea water containing the sea-urchin eggs. The sugar solution is heavier than sea water and is poured so as to avoid mixing in order to have

the sugar solution at the bottom of the tube and a region of transition where the sugar mixes with the sea water so that the density diminishes in the upward direction from the high density of the sugar water to the density of sea water. The sea-urchin eggs, being denser than sea water but less dense than the sugar solution used, sink until they reach a layer whose density is equal to their own, and there they remain suspended even at the highest speeds of revolution of the centrifuge, having no more tendency to precipitate than water droplets suspended in water.

APPENDIX II

Surface Tension

Surface Phenomena

A dead fly or beetle tends to sink to the bottom when immersed in water. This indicates that its average density is larger than that of water. And yet we can see in nature large water beetles running over the surface of a lake just as comfortably as over solid ground seemingly without wetting their feet. Larvae of mosquitoes suspend themselves from the water surface by means of hairlike appendages. But when oil is sprayed on the surface, they sink and suffocate, indicating that their average density exceeds that of water. In fact, under special precautions, a sewing needle or even a heavy metal net may be laid on the surface of water where it will remain suspended as if the water surface were lined with an invisible rubber membrane.

How can these facts be reconciled with Archimedes' principle?

Very simply. They have nothing to do with it. The clue for their understanding is rather provided by the following observations: Raindrops or any blobs of liquid suspended in another fluid are spherical. If we inject, for instance, a small volume of oil into a mixture of water and alcohol whose density is the same as that of the oil, our oil volume will assume a spherical shape. And if we deform it, it will tend to resume the shape of a sphere. For example, setting the oil drop in rotation will flatten it across the axis of rotation (owing to centrifugal d'Alembert forces), illustrating the mechanism responsible for the flattening of the rotating earth, but after rotation ceases, the drop is spherical again.

Why does the blob of liquid tend to remain spherical? Is there any property peculiar to the spherical surface which distinguishes it from all other surfaces? Yes, there is. The sphere has the smallest surface area for a given volume. Any deviation from sphericity involves increase in surface area; resumption of sphericity, a decrease. But why should the fluid particles on the surface of the oil tend to retire into the interior, thus reducing the surface area?

Our First Hypothesis

To answer this question we shall take a step which is novel in our experience. We shall use our imagination and formulate our first hypothesis. In fact, we already took a vague preliminary step in this direction when we referred to fluid "particles." We cannot possibly visualize a fluid (or solid) continuum indefinitely subdivisible. We can visualize motion or changes within a fluid only by imagining it to consist of tiny discrete particles, the motion of which we can follow in our imagination. Thus, owing merely to our habit of thinking and not on the basis of suggestive evidence, we shall ascribe to fluids a granular structure, and we shall give these imaginary granules the tentative name of "molecules." (Having admitted their possible existence in liquids, we have no reason to assume their disappearance when the liquid freezes into a solid.) Adhesion of liquids to solids and observed cohesion forces between parts of the same liquid suggest that the molecules must attract each other (at least at close range).

These assumptions are sufficient to explain why a liquid should always tend to assume a minimal surface. **By "explain" we mean here to provide a visualizable mechanical model of a process.** A molecule m_1 in the interior of a liquid will be attracted by its immediate neighbors uniformly in all directions for symmetry reasons (Fig. 11.14). (We assume the range of attraction to be limited to a submicroscopic distance.) It will thus be in equilibrium. A molecule m_2 on the surface (or near it) will, on the other hand, experience inward forces of attraction from neighbor molecules below it but no upward forces (or negligible ones due to air). Thus, *the molecule will be pulled into the interior of the liquid by the resultant of the molecular forces.* This inward pull tends to remove molecules from the surface into the interior of the liquid, thus explaining the tendency of the liquid to assume the smallest possible surface for a given volume.

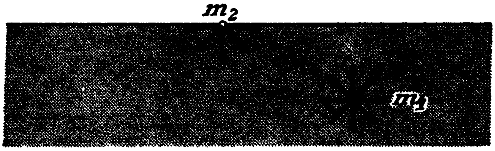


FIG. 11.14. Forces exerted upon interior and surface molecules.

We can also illuminate this situation from a different angle: The molecules on the surface of the liquid which are pulled inward are like weights on a shelf which are pulled earthward by gravity; they possess potential energy. As a molecule enters the interior, it loses its original potential energy. A reduced surface means reduced potential **surface energy.** But as we have seen in Chap. 9 (Appendix), the stable state is one of minimum potential energy. We can therefore state that a blob of liquid tends to assume a shape of minimum surface because the latter represents a shape of minimum potential energy and, hence, a stable configuration.

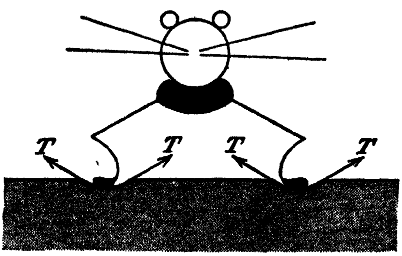


FIG. 11.15. A bug upheld on water surface by surface tension.

Surface Tension

The tendency of a surface to reduce its area lends it properties reminiscent of a rubber membrane. The surface exhibits tension. This **surface tension** explains why insects can walk on a water surface or why a sewing needle may not sink. Figure 11.15 shows a bug on the surface of water. Its weight is neutralized, not by buoyancy (the bug is hardly submerged!), but by the *upward component* of the surface tension forces T which are tangential to the deformed surface.

Surface tension can be demonstrated in a quantitative fashion with a hairpin (Fig. 11.16). When the pin is dipped in a soap solution, a thin film will form which may be bounded at the bottom by a crosspiece C. A small weight W can be kept in suspension by the forces of surface tension exerted by the two surfaces of the film. Thus, **the surface tension can be defined and measured as the force exerted by each surface of the film per unit length of the boundary line C.**

An important difference between the surface tension and the tension of an elastic rubber membrane consists in the fact that the surface tension force in Fig. 11.16 does not increase as the film is distended.

Ingestion of Bacteria

Surface tension forces and considerations of surface energy are of fundamental importance in biology. For example, the important process of phagocytosis (ingestion of dead particles or bacteria by living cells) may be explained by surface-energy considerations. When, for instance, a large mononuclear lymphocyte (blood corpuscle) encounters a staphylococcus (measuring about 1μ, or about one-tenth the size of the blood corpuscle), it will begin to envelop it, and the process finally stops when the coccus has been partly ingested. It has been shown that the process of envelopment involves a loss of surface energy by the ingesting particle, so that phagocytosis can be considered as a change in the direction of a stable configuration of minimum potential energy.

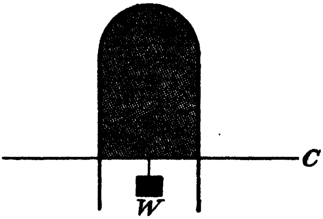
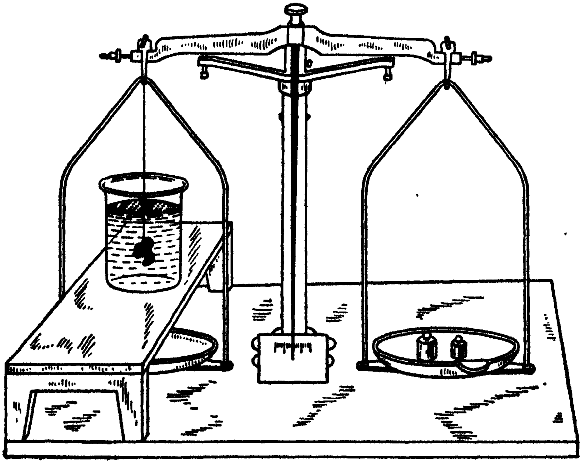


FIG. 11.16. Measurement of surface tension. C is a cross wire which slides easily up and down.

Example 1

You are given an irregular piece of brass and an oil of unknown density. You are asked to determine the density of brass and of the given oil.

Experimental Procedure. If you knew the volume of the brass and its mass, you would have its density $d = m/V$. You find m by weighing and use the scale for an accurate determination of volume as follows: The piece of brass is suspended from a fine thread and submerged in a beaker of water (which we assume for simplicity to have the temperature of 4°C). The beaker is placed on a small table so as not to rest on the scale pan. We thus find the loss of weight suffered by the brass in water, which is equal to the weight of the displaced water. Since at 4°C the weight of water in



grams† is numerically equal to its volume, we have found the volume of the brass specimen and hence its density $d = m/V$.

We now proceed to find the density of oil. By pouring oil into our beaker instead of water, we can find the loss of weight of the piece of brass (whose volume we now know) in oil. This loss of weight must be equal to the weight of the displaced volume of oil. Hence, dividing this loss of weight (*i.e.*, the weight of the displaced oil) by the displaced volume, we find the weight of oil per unit volume, which, in this case, is numerically equal to the mass per unit volume or density.

Results of measurements:

Weight of brass in air: 85 gm.

Weight of brass in water: 75 gm.

Weight of brass in oil: 76.5 gm.

Computation:

Loss of weight in water: $L_w = 85 - 75 = 10$ gm = weight of displaced water.

Hence, volume of brass: $V = 10$ cm^3.

Density of brass: $\underline{d} = \frac{85}{10} = \underline{8.5 \text{ gm/cm}^3}$.

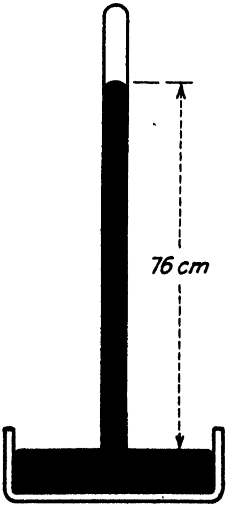
Loss of weight in oil: $L_0 = 85 - 76.5 = 8.5$ gm = weight of displaced oil.

Volume of displaced oil = volume of brass sample = 10 cm^3.

Hence, density of oil: $\underline{d} = 8.5/10 = \underline{0.85 \text{ gm/cm}^3}$.

Example 2

How large is the force in pounds exerted upon a floor of 100 ft^2 owing to the hydrostatic pressure of the atmosphere when the barometer reads 76 cm Hg at 0°C?



The barometer consists of a glass tube about 1 m long filled with mercury. It is inverted and dipped into a dish with mercury. The column in the tube sinks to a height at which its weight is equal to the weight of a column of the atmosphere of equal cross section. The weight of either column per unit cross section is the atmospheric pressure.

---

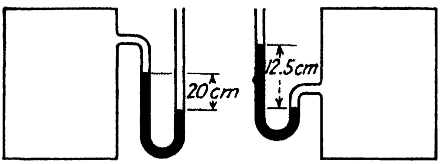
† We define 1 gm weight as the force with which the mass of 1 gm is attracted toward the earth at a specified locality.

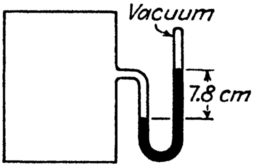
Solution. The weight of a mercury column of 76 cm height and 1 cm² cross section is $p = hdg = 76 \times 13.6 \times 980 = \mathbf{1.013 \times 10^6}$ **dynes/cm²**. Similarly, one can find in British units the weight of a column of this height and a cross section of 1 in.²: $p' = 14.7$ lb/in.² and for the cross section of 1 ft²: $p'' = 14.7(12)^2 = 2{,}120$ lb/ft². Thus, the force exerted upon the floor of area $A = 100$ ft² is

$$\underline{F} = pA = 2{,}120(100) = \underline{2.12 \times 10^5 \text{ lb}}$$

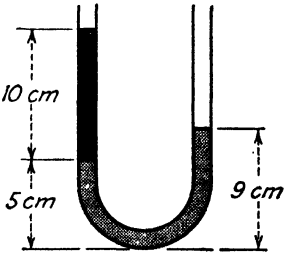
QUESTIONS AND PROBLEMS

1. Find the pressures in these three tanks at an atmospheric pressure of 75.6 cm Hg. The manometers are filled with mercury.





2. A U-tube manometer is filled with two different liquids. Find from the data given in the diagram the ratio of their densities.



3. If a long vertical glass tube is connected to an artery of a dog whose average arterial blood pressure is 100 cm Hg, to what average height will the blood rise approximately (*a*) in air at normal atmospheric pressure? (*b*) in a vacuum?

4. The atmospheric pressure varies with height for the same reason the hydrostatic pressure in a lake varies with depth. If you release a light object, such as a piece of cork, from the bottom of the lake, it will not come to rest before reaching the surface; if, on the other hand, you release a light balloon into the atmosphere from the ground, it will eventually settle in equilibrium at a definite height above ground. Explain the results of these two experiments.

5. It is dangerous in colonic irrigations to raise the enema can too high above the patient, since the hydrostatic pressure in the intestine may become excessively high. Compute the hydrostatic pressure in the intestine for the case when the water level in the enema can is 1 m above the intestine. Express the pressure in grams per square centimeter. Compute the force in kilograms exerted upon an area of 1,000 cm² of the colon. Convert this force into pounds.

6. How deep would a lake have to be for the hydrostatic pressure at its bottom to be equal to atmospheric pressure of 76 cm Hg?

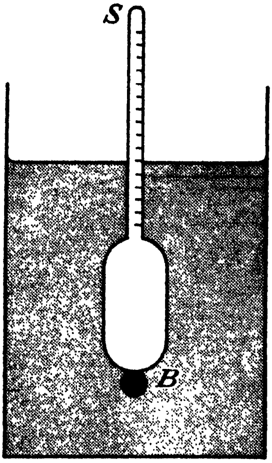
7. How large is the pressure at the bottom of a lake 200 ft deep? Is the pressure exerted upon the fish of this lake subject to variation with variation in atmospheric pressure?

8. It is known that 90 per cent of the volume of floating ice is submerged under water. Find the density of ice.

9. The cross section of a boat is 5 m². Assuming its walls to be vertical, compute how much deeper it will sink when two men each weighing 75 kg get into it.

10. A hydrometer is a device for measuring the density of liquids. The heavy sphere B keeps it in a vertical position, and the uniform thin stem is provided with a uniform scale.

Assume that you prepare a series of solutions dissolving n, $2n$, $3n$, etc., grams of salt per 100 gm of solution. If you then immerse the hydrometer into these solutions, it will sink to different depths and you can make a mark each time on the stem for the position of the water surface. Show that these marks will be equidistant.



11. A 1,300-gm piece of metal weighs 1,150 gm when completely submerged in water. What is its density?

12. A 500-gm wooden cylinder of density 0.5 is to be submerged completely in water by tying a piece of lead to it. How large must that piece of lead be?

13. Is the equilibrium of the cartesian diver stable, unstable, or neutral?

14. The glass diver of a cartesian-diver microbalance weighs 10^{-4} gm in air. How large an air bubble must we introduce into the diver to make it float in water?

15. Compute the centrifugal inertial force per cubic micron which cells of density 1.1 gm/cm^3 experience in a salt solution of density 1.02 gm/cm^3 when they are centrifuged at 60,000 rpm. Assume the cells to be located 5 cm from the axis of revolution.

How does the centrifugal force depend (*a*) on the centripetal acceleration? (*b*) on the difference in densities between suspended particles and solution? (*c*) on mass of the particles? (*d*) How does buoyancy depend on the frequency of revolution of the centrifuge? (*e*) In what way do particles less dense than the surrounding liquid behave differently from particles which are more dense than the liquid?

16. Human blood pressure is measured clinically by placing a "pressure cuff" around the upper arm and inflating the cuff so as to stop the flow of blood through the brachial artery. The air pressure in the cuff is measured by means of a mercury manometer. Then the air pressure is gradually diminished by adjusting an air leak. Explain how you could measure with this equipment (*a*) the systolic blood pressure (the maximum pressure during the heart cycle), (*b*) the diastolic pressure (the minimum pressure during the period of relaxation of the heart), making use of the fact that you can hear the flow of blood through a superficial artery with the aid of a stethoscope. Criticize the accuracy of this method.

17. Work out a simple scheme to measure the instantaneous arterial blood pressure in man.

18. How large is the hydrostatic pressure at the bottom of a jar filled to a depth of 10 cm with mercury when it is (*a*) accelerating upward with $a = 980$ cm/sec^2? (*b*) accelerating downward at the same rate?

BIBLIOGRAPHY

On the microrespirometer:

Lindestrøm-Lang, K., *Nature*, Vol. 140, p. 108, 1937.
Boell, E. J., J. Needham, and V. Rogers, *Proc. Roy. Soc. B*, Vol. 127, p. 322, 1939.

On the microbalance:

Zeuthen, E., *C. R. Lab. Carlsberg*, Vol. 26, p. 244, 1948.

CHAPTER 12

FLUIDS IN MOTION

a. How can we visualize the flow of a "homogeneous" (uniform) fluid?

b. What do streamlines symbolize? What quantitative information do they convey?

c. Is the continuity equation $Q_A v_A = Q_B v_B =$ constant equally applicable to gases as it is to liquids?

d. Are the Newtonian laws of motion applicable to the dynamics of fluids?

e Do you have to assume a molecular (granular) structure of fluids in order to apply Newton's laws of motion to them?

f. Bernoulli's equation can be considered as a special statement of what general law?

g. What energetic interpretation can you associate with the concept of fluid pressure?

h. What is the physical significance of the three terms on each side of Bernoulli's equation (12.3)?

i. As fluid flows through a pipe of *varying cross section* at a constant rate (*i.e.* same number of gallons each minute), do fluid particles suffer acceleration or retardation at any point of the pipe line?

j. Does the pressure vary throughout the above pipe line?

k. Can the pressure in the above example ever become zero? Negative?

l. Do you know a method which measures the fluid velocity directly?

m. Does it require a force to move at a constant speed (1) an ideal, (2) a viscous fluid through a *uniform* pipe?

n. What does the pressure drop along a uniform pipe indicate energetically?

o. What happens to the "lost" mechanical energy in viscous flow?

p. Which is more effective in controlling flow through a pipe, changing its length or its diameter?

q. How does the fluid velocity in viscous flow vary throughout the pipe cross section?

r. How would the velocity distribution in a pipe look in the case of an ideal fluid?

Extension of Newtonian Mechanics to Fluids

Whereas in Chap. 12 we considered fluids at rest (hydrostatics), we shall now turn our attention to the study of fluid motion. In studying the motion of rigid bodies we proceeded in two steps: (1) We worked out a kinematics of rigid bodies which dealt solely with the description of possible motions without reference to the cause of motion, and then (2) we attempted to establish a relation between the actually observable types of motion and the acting forces (dynamics). Accordingly, our next step in extending mechanics to fluids will be the development of (1) a kinematics of fluids, (2) a dynamics of fluids. But here we encounter a familiar difficulty. If you look at a glass pipe through which perfectly clear water (without air bubbles or solid particles) is supposed to flow, you have no way of telling whether the water is standing still, is moving in a straight line, or is full of vortices (so-called *turbulent flow*). You can neither see fluid motion, nor can you imagine it unless you visualize some sort of discontinuity, such as dirt particles suspended in your fluid. You could then follow the path of these tiny solids and assume that the fluid in their immediate vicinity shares their motion. We see the necessity of introducing some kind of particles (real or imaginary ones) in order to be able to visualize fluid motion. As an alternative to imagining suspended particles, we can make use of our molecular hypothesis of Chap. 11. Instead of considering fluids as continua, we shall ascribe to them a granular structure and speak of "fluid particles." Vague as this notion is, it is adequate for our present purpose.

Streamlines

Hydrodynamics is one of the most difficult subjects in physics. We shall confine ourselves to the simplest case of steady flow, *i.e.*, flow where the velocity at any point remains constant.

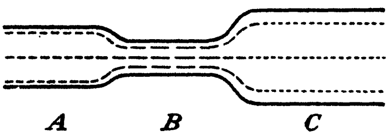


FIG. 12.1. Time exposure of flow in a pipe (or channel) of varying cross section. The streaks of varying length illustrate the path of suspended particles traversed during the exposure time.

Suppose we have a canal and want to know what path fluid particles on the surface pursue in a steady flow and how fast they move at various points. How could we determine it? We could spray the surface with tiny fireflies† at night and take a photograph of the top view. If we choose a very short exposure time (say 1/10,000 sec), we are likely to get a sharp photograph of our fireflies. But if we choose a long exposure time (say $\frac{1}{10}$ sec), every firefly will appear like a streak and we shall have a picture such as Fig. 12.1. Every streak shows the length and direction of the displacement of a firefly at various points in $\frac{1}{10}$ sec. For practical purposes the photograph gives us a distribution of the local velocity vectors over the liquid surface, since the length of a streak is proportional to the local velocity and the direction of the streak gives the direction of the velocity vector.

We can draw through these streaks a series of solid lines which will, at every point, indicate the direction of the local velocity. These lines are called **streamlines.** These streamlines are an invaluable aid in visualizing fluid motion.

---

† This suggestion is not to be taken literally, since living fireflies will hardly oblige us by glowing on the water surface.

Continuity Equation

A closer inspection of Fig. 12.1 shows that the streaks in section B are longer than in sections A and C, and they are shorter in C than they are in A. This means that the fluid flow is fastest in section B and slowest in C. At the same time we notice that B happens to be the narrowest of the cross sections and C the widest. In general, we observe that the larger the cross section the slower the flow!

Why should it be so? Is it always true? What is the precise relationship between flow and cross section?

A simple precise relationship can be derived for a liquid of negligible compressibility flowing through a closed pipe. Imagine that Fig. 12.2 represents a constricting pipe. Suppose the liquid is propelled by pushing the piston P forward. If v_A represents the distance the piston travels in 1 sec (*i.e.*, the fluid velocity at cross section A) and if Q_A is the cross-sectional area at A, then the volume of fluid pushed forward by the piston is equal to $Q_A v_A$. If the fluid is nearly incompressible, the same volume will have to pass through every subsequent cross section of the pipe. The volume passing through every cross section per second is equal to cross-sectional area times velocity. Hence, we can write the following **continuity equation**:

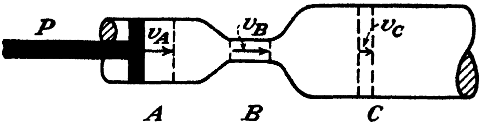


FIG. 12.2. Illustration of the condition of "continuity." The liquid volume injected by the piston into the pipe section A is represented by a cylinder of height v_A. The same volume also passes into the section B (as a cylinder of height v_B) and into the section C (as a cylinder of height v_C). If the injection time is 1 sec v_A, v_B, and v_C represent the fluid velocities in the pipe sections A, B, C.

$$Q_A v_A = Q_B v_B = Q_C v_C = \boxed{Qv = \text{constant}} \qquad (12.1)$$

or

$$\underline{\frac{v_A}{v_B} = \frac{Q_B}{Q_A}}$$

The ratio of the velocities at any two cross sections is equal to the inverse ratio of the cross-sectional areas.

We notice also that the distance between the neighboring streamlines becomes smaller in a region of faster flow (Fig. 12.1). This is a direct consequence of the continuity equation. Suppose Fig. 12.1 represents a horizontal constricting channel. Two neighboring streamlines can be imagined to form the walls of an imaginary "stream channel" (or "streampipe") of rectangular cross section. When the distance between the lines is halved, the cross section of the imaginary channel is halved and, hence, the velocity of flow in that region must be doubled (provided the dimensions of the channel at right angles to the plane of the paper do not change). Thus, we see that the lines of flow, or streamlines, not only represent the direction of the local flow velocity but, through their relative distance, also indicate the relative magnitude of the local velocity. **The closer the streamlines the faster the flow.** We shall encounter symbolic representation of electric and magnetic "fields" by lines of force quite analogous to the above-mentioned representation of **flow fields** (regions of fluid flow) by streamlines.

Bernoulli's Equation

After this kinematic introduction, let us attempt to extend the laws of dynamics, as we developed them for mass points and rigid bodies, to fluid motion. The simplest approach is through the use of the law of conservation of energy. We shall limit ourselves for the time being to the so-called incompressible "ideal fluids," *i.e.*, frictionless liquids.

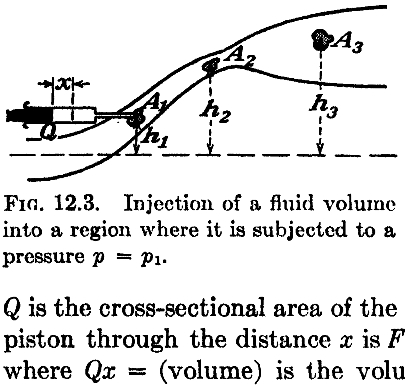


FIG. 12.3. Injection of a fluid volume into a region where it is subjected to a pressure $p = p_1$.

Consider the following concrete situation (Fig. 12.3). We inject a certain volume of fluid through the rubber wall of our pipe system at A_1. We have to do work against the prevailing local pressure p_1 to inject this volume. The force which our piston has to overcome is $F = p_1Q$, where Q is the cross-sectional area of the piston. The work done in moving the piston through the distance x is $F_1x = (p_1Q)x = p_1(Qx) = p_1 \cdot \text{(volume)}$ where $Qx = \text{(volume)}$ is the volume injected with the syringe. What happened to the energy we spent in injecting the fluid? When we withdraw the syringe and allow the same volume to squirt out from our rubber hose, the escaping liquid will possess a kinetic energy equal to the work we have done in injecting it. Thus, we may say that the work $p_1 \cdot \text{(volume)}$ we did was stored in some form of potential energy. Since the gain of potential energy of the injected fluid is associated with an increase in pressure to which it is subjected,† we may call this energy *pressure energy*. For simplicity let us assume that we injected a unit volume: (volume) = 1 cm^3. Then $p_1 \cdot \text{(volume)} = p_1 1 = p_1$, and we can make the following simple statement: **The pressure at any given point in a fluid has the significance of *pressure energy* per unit volume.**

---

† In our example liquid was transferred from a region where it was subjected to a pressure $p = 0$ to a region where it is subjected to a pressure $p = p_1$.

Let us now imagine that we inject the same volume V of fluid at A_2 instead of at A_1. Then, the work we do in injecting it is equal to p_2V or p_2 ergs per unit volume. The potential energy due to pressure will, in general, be different at points A_2 and A_1. The kinetic energy is also different. Owing to the higher velocity in the narrower section at A_2, the fluid element possesses a higher kinetic energy at that point. If the pipe section shown in Fig. 12.3 is in the horizontal plane, the pressure energy and kinetic energy are the only forms of mechanical energy between which transformation can take place. A gain in kinetic energy by the fluid element at A_2 implies a corresponding loss in pressure energy which is given by $\Delta_1 = p_1 - p_2$ for a unit volume. This means that pressure should be lower at A_2 than at A_1. There is thus a gradual drop of pressure in the constricting pipe section between A_1 and A_2. A blob of fluid subject to varying pressure forces acting on its surface is accelerated from the region of higher pressure at A_1 toward the region of lower pressure at A_2.

If the pipe section shown in Fig. 12.3 is in the vertical plane, however, the loss of pressure energy is not the only change in the energy content of a unit volume as it moves from A_1 to A_2. We see that it gains potential energy of position as it moves upward. The potential energy at A_1 (with respect to an arbitrary base line) is equal to $h_1(dg)$ (where dg is the weight of the unit volume) and at A_2 is equal to $h_2(dg)$. Hence, the gain in potential energy is $\Delta_2 = h_2(dg) - h_1(dg)$.

The gain in kinetic energy of the unit volume is given by

$$\Delta_3 = \tfrac{1}{2}\,dv_2^2 - \tfrac{1}{2}\,dv_1^2$$

[where d (the density) is the mass of the unit volume and v_1 and v_2 are the velocities at the points A_1 and A_2, respectively].

Since these are the only energy transformations which take place in the ideal incompressible fluid, the gain in energy $\Delta_2 + \Delta_3$ must be equal to the loss of pressure energy Δ_1: $\Delta_1 = \Delta_2 + \Delta_3$. Substituting the preceding values for Δ_1, Δ_2, and Δ_3, we obtain

$$(p_1 - p_2) = [h_2(dg) - h_1(dg)] + (\tfrac{1}{2}\,dv_2^2 - \tfrac{1}{2}\,dv_1^2) \qquad (12.2)$$

This can also be written as follows:

$$\boxed{p_1 + dgh_1 + \tfrac{1}{2}\,dv_1^2 = p_2 + dgh_2 + \tfrac{1}{2}\,dv_2^2 = \text{constant}} \qquad (12.3)$$

We recognize the meaning of the term

$$\underline{p + (dg)h + \tfrac{1}{2}\,dv^2 = E} \qquad (12.3a)$$

E **is the total mechanical energy of the unit volume** (pressure energy plus positional potential energy plus kinetic energy). Equation (12.3) states that the total mechanical energy of a unit volume at any point A_2 is the same as at any other point A_1; in other words, it remains constant as the unit volume moves through the pipe. This statement is an expression of the law of conservation of mechanical energy as applied to a moving unit volume.

In the case of a horizontal pipe, $h_1 = h_2$, and we get the simplified equation

$$\boxed{p_1 + \tfrac{1}{2}\,dv_1^2 = p_2 + \tfrac{1}{2}\,dv_2^2 = \text{constant}} \tag{12.3b}$$

Equation (12.3) is the famous **Bernoulli equation,** a fundamental equation of motion in fluid mechanics. In this equation p is measured in dynes per square centimeter, d in grams per cubic centimeter, h in centimeters, and v in centimeters per second.

The validity of Bernoulli's equation is not confined to pipes. We could have made the same consideration imagining the pipe walls to be represented by streamlines ("streampipe").

Variation of Pressure with Speed. To examine an interesting consequence of Bernoulli's equation, let us apply equation (12.3b) to the flow through the system of Fig. 12.2. When there is no flow: $v_A = v_B = 0$ and $p_A = p_B$ (p_A and p_B are the pressures in the pipe sections A and B respectively). Let us now see what will happen to p_B if we keep p_A constant but vary the velocity. According to equation (12.3b),

$$p_B = p_A + \tfrac{1}{2}d(v_A^2 - v_B^2)$$

Note that the cross section of B is smaller. Therefore, *under all circumstances* $v_B^2 > v_A^2$. Consequently $(v_A^2 - v_B^2)$ is always negative, and hence $p_B < p_A$. We see thus that **increased velocity is accompanied by decreased pressure! (Bernoulli effect.)**

In fact the pressure at B could even become zero! Namely, when

$$|p_A| = |\tfrac{1}{2}d(v_A^2 - v_B^2)|$$

and negative (tension) when

$$|\tfrac{1}{2}d(v_A^2 - v_B^2)| > |p_A|$$

As we shall see later from the kinetic theory of gases, a negative pressure in a gas is inconceivable. In the case of a liquid, boiling will take place at small positive values of pressure (equal to the vapor pressure of the liquid at the prevailing temperature), thus making the achievement of a negative pressure in a liquid impossible, except perhaps for a short period of time.

Examples of Applications of the Bernoulli Effect. The reduced pressure in regions of rapid flow has many practical uses. For instance, the principle of the aspirator pump, which is widely used in laboratories to create a vacuum down to a few millimeters of mercury of gas pressure† and which also underlies the design of "atomizing sprays" (in which a liquid medication is sucked up in a tube, broken up into a spray of fine droplets, and carried to its destination by an air blast), is based on the Bernoulli effect. Figure 12.4 shows a schematic diagram of an aspirator pump. Water enters at A, acquires a high speed at the constriction C, and leaves the wide exit pipe at B with a reduced velocity. The pressure at the exit point B is atmospheric and, according to the Bernoulli equation, below atmospheric at the point of maximum velocity at C. Hence, if the outlet D is connected to a reservoir which is at atmospheric pressure, air will begin to stream from it toward C.

Instead of water we can pass air through the tubes A and B of the pump, which will function with air as well. If we then connect D to a rubber tubing which dips into a beaker of liquid, the liquid will be "sucked up" into the pump, *i.e.*, pressed in by the excess of atmospheric pressure over the pressure inside the pump. There it will get into the pipe B and emerge in the form of a spray of water droplets. To obtain this effect, the pipe B and the pump enclosure are actually superfluous. Two tubes disposed at right angles to each other as A and D are sufficient.

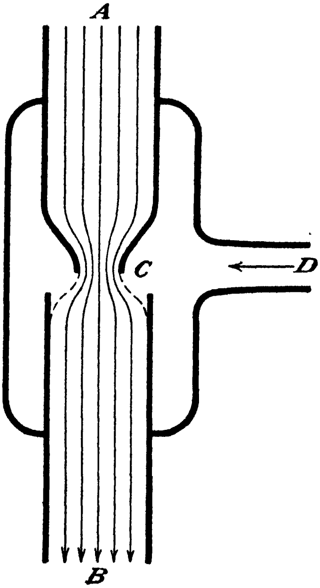


FIG. 12.4.

The Bernoulli effect also explains the following paradoxical experience: When a strong wind is blowing in our face or when we are riding fast in an open car, we have difficulty in breathing. This is due to the Bernoulli effect. The rapid relative motion of air around our nose produces a region of reduced pressure below our nostrils which makes it difficult to catch our breath.

The lift of an airplane or of a bird's wing is also largely due to the Bernoulli effect. A study of streamlines in the flow of air past a wing shows (Fig. 12.5) that the streamlines are more crowded on the upper side of the wing. As we know, the reduced distance between streamlines indicates an increased velocity of flow, and this, in turn, indicates a reduced pressure on the upper side of the wing as compared with the lower side. It is primarily the pressure difference between the upper side and the lower side of the wing which provides the lifting force.[†]

---

† Similar pumps, using a jet of oil or mercury vapor, achieve pressures below 10^{-8} mm Hg.

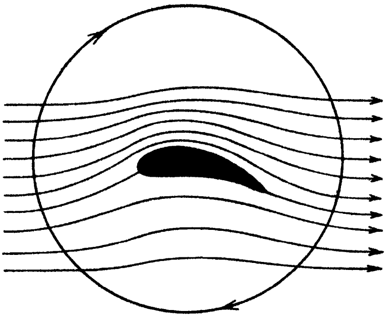


FIG. 12.5. Velocity distribution in air moving past a wing.

In active flight a bird derives its lift from the downward motion of its wings. But many birds make frequent use of a passive mode of flight, the gliding flight, in which they utilize the relative motion of wind past their nearly motionless wings to produce a lift according to the mechanism described above.

Some plant seeds are excellently designed gliders. For instance, the Javan winged seed of *Zanonia macrocarpa* (Fig. 12.6) will pass into a beautiful stable glide when released from any position.

The lift is a force roughly perpendicular to the direction of the moving air (or of the moving wing). It can be shown that only the relative velocity is the essential factor, it being immaterial whether the air moves past the wing or the wing moves past the air.

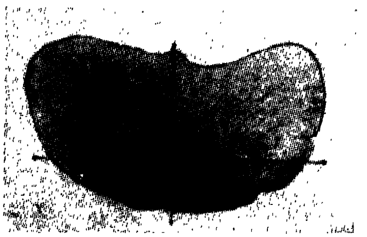


FIG. 12.6. A plant seed (Zanonia) glider. (*From Grimsehl, "A Textbook of Physics," Blackie.*)

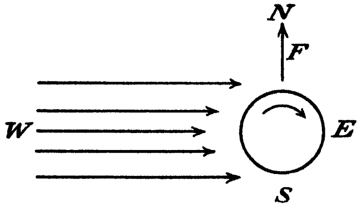


FIG. 12.7. Magnus effect.

A force perpendicular to the direction of motion, similar to the "lift" described above, can also be created without a wing. For instance, if you rotate a cylinder which is exposed to a west wind (Fig. 12.7), air will be accelerated on the north side of the rotating cylinder and slowed down on its south side, resulting in a northward force F (Magnus effect). Such rotating cylinders have been used to replace sails on ships. The same effect is responsible for the peculiar deviation of rotating golf or tennis balls† in flight and for the curved flight of a boomerang.

[†] A more detailed consideration shows that actually a circulation of air is created about the wing, the sense of circulation being clockwise in the upper example. The circulating air moves in the direction of the relative wind velocity (to the right in Fig. 12.5) above the wing and opposite to it (to the left) below the wing. This circulation is the reason why the wind velocity is increased on the upper and decreased on the lower side of the wing.

Example of Deduction from Bernoulli's Principle. We shall now illustrate how Bernoulli's principle can be used to solve special problems in fluid mechanics.

How fast does a liquid escape from a container?

We can apply Bernoulli's principle to this problem by considering the bottle as a very wide pipe of diameter D_A (cross-sectional area $A = \frac{1}{4}\pi D_A^2$) and the exit hole as a constriction in that pipe of cross-sectional area $a = \frac{1}{4}\pi D_B^2$ (Fig. 12.8).

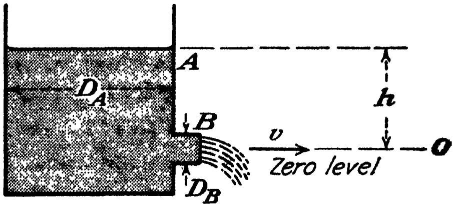


FIG. 12.8. Outflow from a tank.

Applying equation (12.3) to this case, considering the liquid surface A as our cross section 1 and the opening B as our cross section 2, we get

$$p_A + dgh + \tfrac{1}{2}\,dv_A^2 = p_B + dg \times 0 + \tfrac{1}{2}\,dv_B^2$$

The following considerations simplify this equation: (1) $p_A = p_B =$ atmospheric pressure, since at both levels A and B the liquid is in direct contact with the atmosphere and is subjected to its pressure. Hence, p_A and p_B cancel out in the above equation. (2) Since the cross section at A is very much larger than at B, the velocity of flow v_A is negligibly small as compared with v_B, and there is even more justification to neglect v_A^2 against v_B^2. Thus, we omit the last term on the left side of our equation. (3) Finally, we can choose arbitrarily our zero level from which h is to be measured. We choose it so as to make $h_B = 0$.

Our equation is thus reduced to

$$dgh = \tfrac{1}{2}\,dv_B^2$$

from which follows

$$\underline{v_B = \sqrt{2gh}}$$

We see that this velocity is independent of the density of the liquid.

---

† One speaks in the terminology of tennis of a "chop" and in golf terminology of a "slice."

It is also interesting to note that v_B, the "velocity of efflux," is such as if the liquid particles had fallen freely from the height of A to the level of B. For a particle falling freely from A to B we compute

$$(mg)h = \tfrac{1}{2}mv^2$$

and hence

$$\underline{v = \sqrt{2gh}}$$

as above.

Measurement of Flow. The most widely used methods of measurement of fluid velocities are based on Bernoulli's equation, such as the Venturi meter, shown schematically in Fig. 12.9.

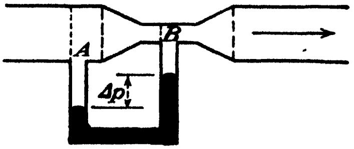


FIG. 12.9. Principle of the Venturi meter.

Let us set up the Bernoulli equation between two levels A and B for the constricted pipe of Fig. 12.9:

$$p_A + \tfrac{1}{2}\,dv_A^2 = p_B + \tfrac{1}{2}\,dv_B^2$$

The continuity equation in the above case is

$$Q_Av_A = Q_Bv_B$$

or

$$v_A = \frac{Q_B}{Q_A}\,v_B$$

Substituting this value of v_A in Bernoulli's equation, we get

$$p_B + \tfrac{1}{2}\,dv_B^2 = p_A + \tfrac{1}{2}d\left(\frac{Q_B}{Q_A}\right)^2 v_B^2$$

rearranging terms and factoring our v_B^2, we obtain

$$v_B^2\left[1 - \left(\frac{Q_B}{Q_A}\right)^2\right] = \frac{2}{d}\,(p_A - p_B) = \frac{2}{d}\,\Delta p$$

Solving for v_B,

$$v_B = \sqrt{\frac{2\,\Delta p}{d[1 - (Q_B/Q_A)^2]}}$$

Since Q_A and Q_B are the constant dimensions of the instrument pipe and d the density of the flowing fluid, we see that *we merely have to measure the difference* Δp between the pressures in the wide and constricted portions *to obtain the velocity of flow through our pipe.* The volume flowing per second is volume $= v_BQ_B$ (where Q_B is the cross-sectional area of the pipe section B).

The Problem of Measuring Blood Flow

Could the Venturi meter be used for the determination of blood flow in animal experiments? In principle, yes. But there are serious objections to this method of measuring blood flow.

1. The blood vessel must be exposed and cut. This requires anesthetizing the animal and maintaining it under anesthesia. The response of an anesthetized animal to drugs and physiological stimuli differs greatly from the normal state. (For instance, some poisons don't affect the anesthetized animal.)
2. The contact of blood with the solid walls of the instrument makes it clot. Hence, a drug (such as heparin) has to be administered to the animal to prevent blood clotting. But then in addition to anesthesia, the responses of the animal are modified by the anticoagulant.
3. The flow of blood in the arteries is not uniform but is a rapidly pulsating one. The usual mercury manometer which measures Δp is too sluggish to follow the pressure variations accurately. The instantaneous velocity thus cannot be obtained with precision, and the average velocity cannot be estimated simply.
4. The constriction at B introduces an additional "resistance" to flow which is absent normally.

We shall discuss in Chap. 24 an electromagnetic method of blood-flow measurements which avoids these difficulties.

Generalized Bernoulli Equation for "Real" Fluids

If we apply the Bernoulli equation, $p_1 + \frac{1}{2} dv_1^2 = p_2 + \frac{1}{2} dv_2^2 = E$, to the flow in a pipe of uniform cross section ($Q_1 = Q_2$ and hence $v_1 = v_2$), we shall expect from the above equation to find $p_1 = p_2$. In other words, the pressure in a uniform pipe should remain constant. This, however, contradicts actual experience. Figure 12.10 shows two side tubes 1 and 2 which act as manometers attached to a uniform horizontal pipe. We find that the pressure diminishes as we proceed downstream. What does this pressure loss mean? As we recall, the pressure p at a point has the significance of the "pressure energy" of a unit volume. Hence, a loss of pressure indicates an energy loss taking place in the moving fluid. (The kinetic energy must be the same at 1 as it is at 2 because a constant cross section secures a constant velocity.) The energy of a unit volume at 1 must exceed the value at 2 by an amount which we shall call p_L (pressure loss). (Such a loss p_L must also exist in a nonuniform pipe.) Equating p_L to the change in the sum of the pressure energy plus the kinetic energy, we get

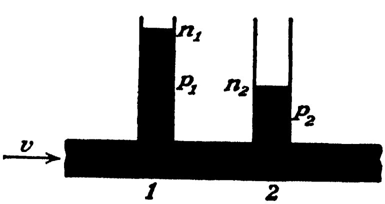


FIG. 12.10. Pressure drop in viscous flow along a uniform pipe.

$$p_L = (p_1 + \tfrac{1}{2} dv_1^2) - (p_2 + \tfrac{1}{2} dv_2^2)$$

or rewriting,

$$\boxed{p_1 + \tfrac{1}{2} dv_1^2 = p_2 + \tfrac{1}{2} dv_2^2 + p_L} \tag{12.4}$$

This is a **generalized Bernoulli equation** which can be used with fluids encountered in practice ("real fluids" or "viscous fluids").

The energy loss of the moving fluid is vaguely analogous to the energy dissipation of a package sliding along the floor. We ascribe the energy loss in both cases to "friction." All fluids exhibit internal friction, or **viscosity,** in varying degree. Molasses is more viscous than alcohol. The more viscous the fluid the higher is p_L in equation (12.4).

We can also interpret the above equation by saying: We must maintain the pressure difference p_L to make the viscous fluid move at a constant rate from point 1 to point 2.

Poiseuille's Law

How does the rate of flow depend on the pressure drop, the dimensions of the pipe, and the nature of the fluid? This question was answered by a French physician, J. L. M. Poiseuille, who was interested in the flow of blood through the capillary blood vessels and decided that he could not hope to understand the complicated process in the living animal before comprehending the simpler process in an artificial circulation model in vitro. Poiseuille found the following law (for "laminar flow"†):

$$q = \frac{\pi R^4}{8\eta} \frac{\Delta p}{L} \tag{12.5}$$

In this equation the symbols have the following meaning: q, the volume discharged through a pipe per unit time; R, the pipe radius; L, its length; Δp, the pressure difference across the pipe; η, the so-called **coefficient of viscosity** which measures the internal friction of the fluid. We shall discuss the physical meaning of η below.

We see from equation (12.5) that, whereas the discharge q is merely inversely proportional to the length of the tube, it is directly proportional to the *fourth power* of the tube radius, so that narrowing the diameter to $\frac{1}{3}$ of the original size will reduce the flow to $(\frac{1}{3})^4 = \frac{1}{81}$ of its initial magnitude! The circulation of blood through our tissues is controlled by cross-sectional changes of the capillary blood vessels. We see that only minute changes in radius are required to produce a marked effect on blood flow.

For future reference we shall rewrite equation (12.5) as follows:

$$q = \frac{R^2\,\Delta p}{8\eta(L/\pi R^2)} = \frac{\Delta p \pi R^2}{8\pi\eta(L/\pi R^2)}$$

[$\pi R^2 = A$ is the cross-sectional area of the pipe. $\Delta p = p_1 - p_2$ and $\Delta p(\pi R^2) = p_1 A - p_2 A = F_1 - F_2 = \Delta F$, where F_1 and F_2 are the forces exerted on the fluid cross section at points P_1 and P_2 where the pressures are p_1 and p_2 and ΔF is the unbalanced force acting on the liquid column extending from P_1 to P_2.] We can write

---

† See pp. 163, 164 for description of laminar flow as distinguished from turbulent flow.

$$\boxed{q = \frac{\Delta F}{8\pi\eta(L/A)} = \frac{\Delta F}{\mathcal{E}(L/A)}} \tag{12.5a}$$

This puts **Poiseuille's law** in a form analogous to **Ohm's law** for the flow of electrical currents which we shall consider later. $\mathcal{E} = 8\pi\eta$ is a specific property of the viscous fluid, measuring its viscosity.

The coefficient of viscosity η is defined as follows: Imagine the smooth plane bottom of an infinitely wide and long river. Visualize a viscous liquid flowing so that the velocity is zero at the bottom and increases linearly with the height y above the bottom. The rate of change of velocity v with the height (the velocity "gradient") is $\Delta v/\Delta y$. The fluid tends to drag the bottom in the direction of flow. The force of drag exerted upon each unit area of the bottom is expressed by F/A. Now η is defined as the ratio: $\dfrac{F/A}{\Delta v/\Delta y}\cdot$ In words: It is the drag per unit area in a flow of unit gradient, $\left(i.e., \text{ where } \Delta v/\Delta y = 1\ \dfrac{\text{cm/sec}}{\text{cm}}\right)$, of the given fluid. The unit of η is 1 dyne sec/cm^2. It is called the *poise* after Poiseuille.

Velocity Distribution in Laminar Flow

Poiseuille and his predecessors observed rather peculiar effects when they watched blood corpuscles move through the fine capillaries of living frogs. At times two neighboring corpuscles seemed to move equally fast, then one of them or both would occasionally change the speed of motion. Often they were seen to rotate, so that the erroneous idea was accepted originally that they possessed the ability to move by themselves. Poiseuille, however, found that it is not so. The explanation for these observations is provided by the mechanism of *laminar* flow.† When a viscous fluid moves through a pipe (such as a blood vessel), the velocity is not uniform throughout the pipe. Next to the wall the frictional retardation of the fluid is greatest, and the fluid particles are at rest in immediate contact with the wall. The farther from the wall the particles are the faster they move. The greatest velocity is found at the center of the pipe. We can visualize these conditions by imagining a row of particles suddenly introduced into the fluid along a diameter AA

---

† In this type of flow, fluid layers, or lamina, slide past each other. This flow is distinguished from *turbulent* flow, in which vortex motion exists. Laminar flow changes into turbulent flow under given conditions when the fluid velocity is increased beyond a critical value. Poiseuille's law does not hold for turbulent flow.

(Fig. 12.11). Taking a snapshot 1 sec later, we shall find these particles elsewhere. They are found to lie now along a curve (a parabola). Each of them has traveled in this second a distance equal to the magnitude of its velocity. This parabola thus gives us a picture of the "velocity distribution" in the pipe.

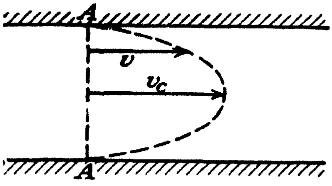


FIG. 12.11. Velocity distribution in laminar pipe flow.

We can now explain the above-mentioned observations made on the blood corpuscles in capillaries. Particles located symmetrically to the center of the pipe move about equally fast. If a particle wanders off center, it will get into the slower layers ("lamina") of the moving fluid and will be seen to slow down. Particles close to the wall will tend to rotate owing to the difference in velocity of the fluid, which moves more slowly on the side of the particle nearer the wall than it does on the side of the particle nearer the pipe center.

APPENDIX

On the Concept of "Pressure Energy"

1. The work done on a unit volume when it moves in a **pressure field** (region of varying pressure) from a point where the pressure is p_1 to a point where it is p_2 is given by $(p_1 - p_2)$. If the pressure p_2 happens to be equal to zero, the work done on the unit volume as it moves from a region where it is subjected to pressure p_1 to a region where the pressure is zero is equal to p_1. *Thus, the pressure to which a fluid element is subjected in a pressure field measures the amount of energy which the fluid element possesses* **relative to a point of zero pressure** by virtue of its location in the region of pressure p. This definition of the "pressure potential energy" at a point is analogous to the definition of gravitational potential energy at a point which is, by definition, equal to the kinetic energy gained by a mass as it moves from the given point to some specified position, called the reference level.

2. In Bernoulli's equation,

$$p_1 + dgh_1 + \tfrac{1}{2}\, dv_1^2 = p_2 + dgh_2 + \tfrac{1}{2}\, dv_2^2 \qquad (1)$$

we can assign an energetic interpretation to each of the terms as follows:

a. $\frac{1}{2}\, dv^2$ measures the amount of work the unit volume can do when its velocity is allowed to drop from the value of v to zero (frictional losses being excluded).

b. hdg measures the amount of work the unit volume can do by descending from the height h to the height zero *while moving in a vacuum* (in the fluid its apparent weight is zero).

c. p measures the amount of work the unit volume can do in moving from the region of pressure p to a region of pressure zero in a pressure field.

3. The following example illustrates the energy transformations in a circulating system:

First, let us visualize a pressure field. In the tank of Fig. 12.12 we have a pressure field. Owing to gravity the pressure increases with depth. We have shown that a fluid element experiences a buoyant force in such a pressure field. This force is the resultant of all the hydrostatic-pressure forces exerted upon the surface of the volume element.

A similar pressure field exists in a constriction which joins two pipes of unequal diameter through which a fluid is flowing (Fig. 12.13). The pressure p_A is uniform in the wide pipe, and the lower pressure p_C is constant in the narrow pipe. In the section *B* the pressure decreases from p_A to p_C. If we imagine a fluid element in this cone, its surface will be subjected to forces similar to the pressure forces which accounted for buoyancy in the previous example. In this case the resultant of the pressure forces points toward the pipe *C*, and these forces accelerate our fluid element in the direction from *A* toward *C*. The gain in kinetic energy ($\frac{1}{2}\,dv_C^2 - \frac{1}{2}\,dv_A^2$) is equal to the loss of "pressure energy" ($p_A - p_C$) of our unit volume in the horizontal pipe.

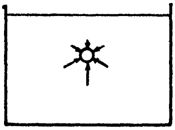


Fig. 12.12. A liquid element in a pressure field.

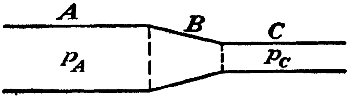


Fig. 12.13.

Let us now consider a circulation of an ideal liquid in a horizontal pipe (no change in gravitational potential energy) (Fig. 12.14).

The frictionless fluid moves forever in a steady flow in our closed system. We have the highest speed and the lowest pressure (p_A) in the uniform pipe *A*. As the volume element enters *B*, it is decelerated by the existing pressure field. In *C* it has a smaller kinetic energy than before but a higher pressure energy (p_C); *i.e.*, it can now do work or gain kinetic energy by moving from the region of pressure p_C to the region of lower pressure p_A. As it enters *D*, it is accelerated in the pressure field. As it gains kinetic energy, it loses pressure energy. At all times the sum of pressure energy plus kinetic energy of a unit volume remains constant. This is expressed by Bernoulli's equation:†

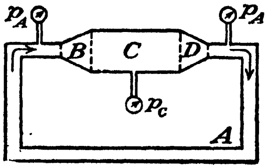


Fig. 12.14.

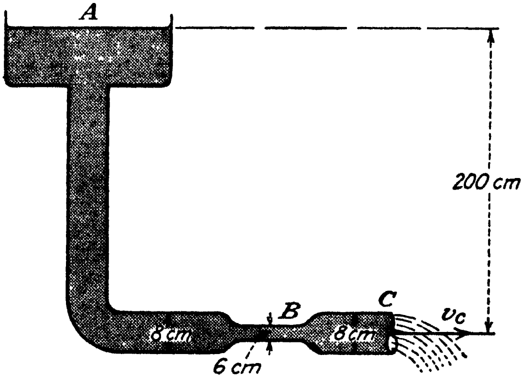
$$p_A + \tfrac{1}{2}\,dv_A^2 = p_B + \tfrac{1}{2}\,dv_B^2 = p_C + \tfrac{1}{2}\,dv_C^2 = p_D + \tfrac{1}{2}\,dv_D^2 \cdots$$

Example

A salt solution of density 1.10 gm/cm^3 is flowing in the illustrated system. Find (1) the velocity of outflow in centimeters per second, (2) the rate of discharge *Q* in cubic centimeters per second, (3) the velocity in section *B*, (4) the pressure in section *B*.

---

† p_B, v_B and p_D, v_D refer to a specified point in the constricted sections *B* and *D*.



Solution. 1. If we neglect friction, the answer to our first question is given by the solution of the example illustrated by Fig. 12.8. Since we did not make any special assumptions about the length and shape of the outflow tube ending at B, we can transfer the solution obtained in that case to our present case by writing C instead of B. We obtain $\underline{v_C} = \sqrt{2gh_A} = \sqrt{2 \times 980(200)} = \underline{626\text{ cm/sec}}$.

2. $\underline{Q} = (\pi R_C^2)v_C = \pi(4^2) \times 626 = \underline{31{,}450\text{ cm}^3\text{/sec}}$.

3. Continuity equation: $(\pi R_B^2)v_B = (\pi R_C^2)v_C$

hence,
$$\underline{v_B} = \frac{\pi R_C^2}{\pi R_B^2} v_C = \left(\frac{4}{3}\right)^2 \times 626 = \underline{1{,}113\text{ cm/sec}}$$

4. Bernoulli's equation between sections B and C,

$$p_B + \tfrac{1}{2} dv_B^2 = p_C + \tfrac{1}{2} dv_C^2$$

If we arbitrarily set $p_C = 0$, the value of the p_B we shall get will have the meaning of an excess of p_B over atmospheric pressure (or a deficit). Thus we write

$$\underline{p_B} = \frac{1}{2} d(v_C^2 - v_B^2) = \frac{1.1}{2} [(626)^2 - (1{,}113)^2] = \underline{-4.66 \times 10^5\text{ dynes/cm}^2}$$

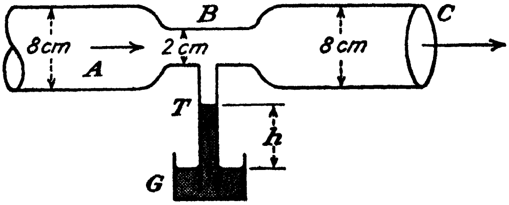
The negative sign indicates a pressure deficit, *i.e.*, the pressure at B is 4.66×10^5 dynes/cm² below the atmospheric pressure whose normal value is 1.013×10^6 dynes/cm².

QUESTIONS AND PROBLEMS

1. Water flows through a cooling section A which has the cross section of a rectangle 2 by 10 cm and is 80 cm long in the direction of flow. From there the flow proceeds through a pipe B of 2 cm diameter. If the discharge is 200 cm³/sec, find the linear velocity in sections A and B.

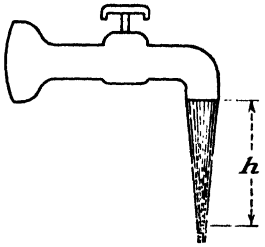
2. Imagine Fig. 12.1 to be the top-view cross section of a horizontal open channel through which water flows so that its velocity does not appreciably vary with depth. Streamlines are made visible by injection of ink at suitable points of the surface. One finds that the distance between two neighboring streamlines varies as follows from section to section: A, 2.1 cm; B, 0.7 cm; C, 2.8 cm. How does the velocity vary from section to section?

★3. How fast must air flow through pipe section A in order to suck up a mercury column of $h = 6$ cm from the beaker G into the vertical glass tube T. The air escapes into the atmosphere at C. (The density of air at standard conditions is 1.29 gm/liter.)



4. Water flows through a pipe of the dimensions shown in Prob. 3. It escapes into the atmosphere at C with the velocity $v_C = 10$ cm/sec. What is the pressure in section B in centimeters of mercury if the atmospheric pressure is 75.0 cm Hg?

★5. Water escapes from a faucet and forms a jet of diminishing cross section as shown in the figure. If the water leaves the faucet with the velocity of 6 cm/sec, (a) how large is its velocity $h = 10$ cm below the faucet opening? (b) how large is the diameter of the jet 10 cm below the faucet as compared with the diameter of the opening of the faucet?



6. Design a Venturi meter of convenient dimensions (Fig. 12.9) and appropriate manometric fluid for the measurement of blood flow in the carotid artery of an anesthetized and heparinized dog. Consult a physiology text to get additional data which you consider pertinent.

7. Plot a graph of manometer reading versus rate of flow for the Venturi meter shown in Fig. 12.9. Assume the diameter of section B to be one-fourth of that of section A. Discuss the practical limitations of this instrument which you can foresee. How does its sensitivity depend on the density of the fluid? on its velocity? on the constriction ratio?

8. Blood serum is flowing through a uniform glass pipe with a speed of 10 cm/sec. The pressure drop between two manometers placed 30 cm apart is 20 cm Hg. How much mechanical energy is converted into heat owing to fluid friction when 1 liter of serum passes through the pipe?

9. What happens to the kinetic energy which the heart imparts to the blood?

10. Blood is flowing through a capillary blood vessel 20μ (2×10^{-3} cm) in diameter and 0.5 mm in length at the average speed of 0.5 mm/sec. How large is the pressure difference between the ends of the capillary (a) in dynes per square centimeter, (b) as expressed in centimeters of water? Assume that the density of the blood is 1.04 gm/cm^3 and its viscosity is 0.072 poise.

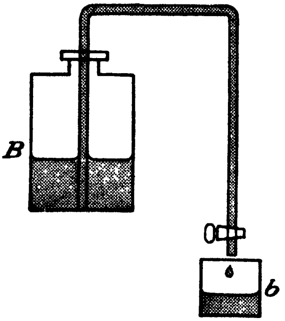
11. How will the blood flow through a capillary blood vessel change when its diameter contracts to one-fourth of its normal size? (Assume that its length and the pressure drop across the capillary do not vary.)

12. Does the blood continue to flow between two successive contractions of the heart? How could it be maintained in motion during that period?

13. Can you design a circulating pump (a crude model of a heart) using a rubber bulb and self-made rubber valves? (Make a diagram.)

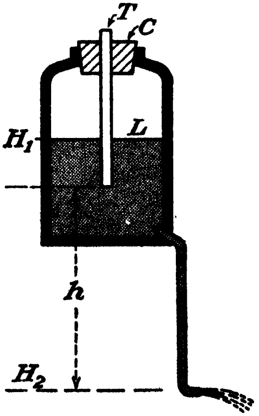
14. Could you estimate to what height blood may squirt out from an injured artery?

15. The figure shows a **siphon,** a device frequently used in medical laboratories and in surgical practice. Our diagram shows a siphon tube siphoning physiological saline solution from bottle *B* into beaker *b*. Rubber-tube siphons are also used to drain cavities (for instance, the abdominal cavity) after operations. Wet gauze can be used as a siphon in an emergency.



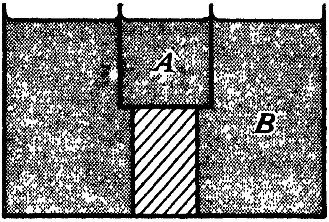
Explain the action of the siphon. Under what conditions will the siphon fail to work? What factors will affect the rate of discharge of the liquid? Will the rate of flow remain constant as liquid escapes from *B*?

★16. The drawing illustrates a Mariotte bottle, which is frequently used in biological laboratories to obtain a constant flow of a liquid. Without the open tube *T* and



cork C, the outflow of the liquid would slow down as the liquid level L sinks. The tube T and airtight cork C, however, make it possible to get a constant outflow despite the sinking of the liquid level L. The fluid flows as if it were driven out by a pressure difference corresponding to the liquid height h. Explain the action of the Mariotte bottle.

17. How could you siphon the liquid from container A into container B?



CHAPTER 13

FIELDS OF FORCE: 1. THE GRAVITATIONAL FIELD

a. What is a velocity field?

b. What is a vector field?

c. What data about the local flow can be determined from the streamline diagram (Fig. 13.1)? How can they be determined?

d. From what premises did Newton induce his law of gravitation?

e. How was the gravitational constant determined?

f. What is the difference between the gravitational constant G and the acceleration due to gravity g?

g. Can buoyancy be considered as a gravitational effect?

h. How do you define the gravitational field intensity?

i. How could you measure the strength of the gravitational field at a point?

j. Is field intensity a scalar?

k. If a body has negative potential energy with respect to a certain level, does it mean that this body cannot do any work? Explain.

l. Why is no work required to move a body along an equipotential line?

m. How do you determine the lines of force when the equipotential lines are given?

n. How can you find the equipotential lines when the lines of force are given?

Vector Fields

In Chap. 12 we arrived at a new concept, the concept of a **vector field,** without actually labeling it as such. A vector field is a region in space to each point of which we can assign a definite vector. For instance, suppose we have a source and a sink in a water tank (Fig. 13.1). As water enters the tank from the source (+), it does not flow along the shortest path toward the sink (−) but spreads out and follows streamlines which can be determined by our "streak method" illustrated in Figure 12.1. The streamlines and streaks so found (Fig. 13.1)† assign to every point in the tank a fluid velocity vector of a definite magnitude and direction. We have thus described a vector field wherein a velocity vector is assigned to every point in our region of flow, or **"field" of flow.** The method of streamline representation offers a simple means of symbolic description of such vector fields. The direction of the vectors is given by the tangent to the **streamlines, or lines of flow,** and the variation in the magnitude of the vector from point to point in the field is determined by the variation in the separation between the streamlines (see Chap. 12), the velocity being highest where the flow lines are closest together (in Fig. 13.1 at the sink and the source) and lowest where the streamlines are farthest apart. For instance, if we adopt the convention of indicating a local flow of 1 $cm^3/(cm^2)(sec)$ by drawing one flow line through each square centimeter (the area being perpendicular to the flow), then by counting the density of flow lines (*i.e.*, the number of lines per square centimeter), we shall be able to determine the local "flux" (*i.e.*, the rate of flow) at any point in the flow field.

---

† This diagram must be pictured in three dimensions. The streamlines are symmetrical with reference to the axis passing through the sink and the source points.

The method of symbolic representation described above is not limited to velocity fields. It can be generalized and applied to all kinds of "vector fields." We shall make use of this method in connection with fields of force which will be introduced in this chapter and used extensively in subsequent parts of this text.

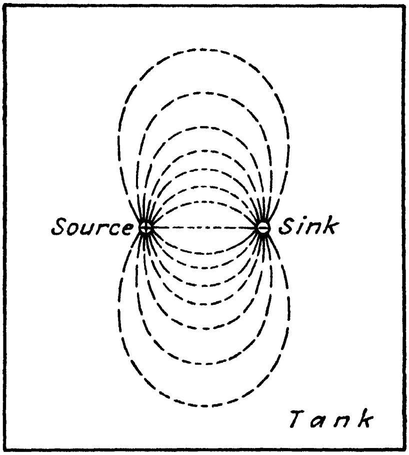


FIG. 13.1. Streamlines in a tank in which water enters from a source and flows out through a sink.

The Gravitational Field

We could (at least in our imagination) perform experiments in the vicinity of our planet which would allow us to map in its surroundings a vector field in the same fashion that we can map a field of flow by means of the streak method in the surroundings of a source or a sink. We could release little stones at various heights above the ground, ranging from a few feet to hundreds of miles. We could release them over America, Europe, Australia, or the North Pole. We should observe that all of them fall aiming toward a common point of intersection, the center of the earth, just as dust particles in a field of flow move toward a sink which sucks them in. By means of a ruler and stop watch, we could measure the acceleration of the falling bodies at all these points and thus assign a definite acceleration vector to every point. It would be a radial acceleration field. The vectors would aim at the same point everywhere, and if we draw **field lines** (analogous to the streamlines of the flow examples) radially toward this point of intersection, we shall obtain field lines which diverge as we recede from the center. In the field representation of flow, we saw that a separation of streamlines meant a diminution of the velocity of flow. By analogy we now suspect that the divergence of our "lines of acceleration" means a diminution of acceleration with the distance from the center of the earth.

Since force is mass times acceleration, we also suspect that the weight of a body would also diminish with the distance from the earth's center. In fact, it would be even simpler to map a "force field" about the earth than an "acceleration field." We could use a spring scale with an attached unit mass. The scale could serve at the same time as a plumb line, giving the direction of the force, and as a force meter, indicating the variation of its magnitude from point to point. The force field we should thus establish would be the **gravitational field** of the earth. We could call the force exerted on the unit mass at a point the **local field intensity.**

Planetary Motion

The path which actually led to the discovery of the laws of gravity was not so simple and straightforward as our imaginary experiments of the preceding section. They were inferred from painstaking observations through subtle reasoning by one of the most powerful minds of all times.

The observation of the motions of the planets can be considered as mankind's most important step toward the creation of science. The observed regularities and the repetitive nature of astronomical phenomena, which made possible prediction of such phenomena as the eclipses of the sun and the moon, suggested an analogy to man-made mechanisms and stimulated speculation aiming to unravel the principles of action of the "celestial machine." Accurate knowledge of the motion of the planets (which are not subject to so many disturbing influences as the bodies we observe in our daily life) proved to be a most fruitful store of factual material for the inference of the basic laws of nature as well as for the test of predictions based on these laws.

One of the outstanding observers of planetary motions was Tycho Brahe, who amassed a tremendous amount of data surpassing in accuracy previous work. The analysis of these data enabled his great associate and successor Kepler to discover the simple approximate laws of planetary motion which bear his name.

After a painstaking analysis of Tycho's data, in the course of which he tried 19 different assumptions about the shape of the planetary orbits, Kepler finally found that the ellipse fitted Tycho's data very well. He did his first work on the planet Mars and eventually generalized his findings for all planets.

Kepler's laws of planetary motion are as follows:

1. Each planet describes an ellipse with the sun at one of its foci (F_1 in Fig. 13.2).

2. The radius vector drawn from the sun to a planet sweeps out equal areas in equal times (see shaded areas in Fig. 13.2).

3. The squares of the periods T of revolution of the planets are proportional to the cubes of their mean distance R from the sun.

$$\frac{T_1^2}{T_2^2} = \frac{R_1^3}{R_2^3}$$

Kepler's laws are not exact but merely good approximations. It is fortunate that Tycho's observations, which Kepler used, were too crude to reveal the fine details of planetary motions, since they might have obscured the simplicity of the law of gravitation and might have delayed its discovery.

Kepler's laws give the impression that the sun stands still and that the planets are attracted only by the sun. Actually, each planet is also attracted by its neighbor planets, which leads to so-called *perturbations* of its motion. It was the study of such perturbations of known planets that led to the discovery of the planets Neptune and Pluto.

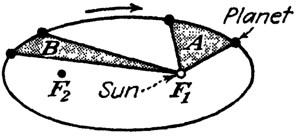


FIG. 13.2. Illustration of Kepler's law of areas. Area A is equal to area B.

The assumption of a stationary sun is also wrong. Whenever we have two planets or stars forming a binary system, *both of them revolve* about their common mass center. The centripetal acceleration of both planets is a consequence of their mutual gravitational attraction and is given by the familiar expression for the centripetal acceleration: $a = \omega^2 R$ (where R is the distance from the mass center of the two bodies to the mass center of the body considered).

Newton's Law of Universal Gravitation

It occurred to a number of people that an attractive force between the sun and the planets might be responsible for the curvature of the path of the planets. Kepler himself assumed the earth and the sun to be two large magnets attracting each other. But such speculations were fruitless, since they yielded no new knowledge. Newton realized that the criterion of a justifiable scientific speculation is its fruitfulness in leading to new knowledge. **To say that a scientific hypothesis is correct means no more than to say that the consequences deduced from it are confirmed by experiment.** To him, to explain planetary motions meant to adopt a set of fundamental assumptions as the basic laws of nature and to deduce from them laws which, if the assumed laws were "correct," should be identical with the observed laws. He had part of the required set of laws in his three basic laws of motion. He needed an additional assumption, which occurred to him, according to a legend, as he saw a falling apple. His inductive reasoning might have run about as follows: "This body is attracted toward the earth. If I raise it to a greater height, it still falls toward the earth. If I remove it as far as the moon's orbit, it still would be attracted toward the earth, though less strongly. Shouldn't the moon itself then also be attracted and be falling toward the earth? If the moon and earth attract each other, shouldn't the same relationship also exist between Jupiter and its satellites, the sun and its planets? But if any two celestial bodies attract each other, is not the resultant force of attraction due to the sum of the attractions which individual small particles of each body exert upon the particles of the other? Does not a stone on the earth attract a stone on the moon? And if it does, shouldn't then a stone on the earth attract another stone on the earth?"

Newton induced the generalization that **every mass in this universe attracts another mass.** He went far beyond his experience in making this generalization, since he observed evidence of mutual attraction of only a few celestial bodies. Two things still remained to be done. (1) A quantitative expression had to be found determining the force of interaction as a function of the masses of the interacting bodies, their distance from each other, and possible other factors which might influence the force. (2) The universal validity of that expression had to be checked by experiments in the laboratory and by applying it to planetary motion.

Newton assumed that the force depends only on the masses and on their distance and on no other factor, such as, for instance, the surrounding medium. The gravitational force between two stones under water is the same as under otherwise equal conditions in a vacuum.† Newton inferred the mathematical form of the law from Kepler's laws of planetary motion. This is another example of the process of induction. Kepler's laws are statements restricted in their validity to a few objects in this universe, whereas the law of gravitation which Newton induced from Kepler's laws is universally valid for all matter in the universe.

Let us see how the dependence of the force upon the distance can be inferred from Kepler's laws. Consider the motion of two planets about which Kepler's third law makes the following statement:

$$\frac{T_1^2}{T_2^2} = \frac{R_1^3}{R_2^3} \qquad \text{or} \qquad T_1^2 = \left(\frac{T_2^2}{R_2^3}\right) R_1^3 \tag{1}$$

Let us now assume that the force which imparts the centripetal acceleration to the planets is the gravitational attraction of the sun and that

---

† One must not forget, however, that under water, in addition to gravity, there will be a force of buoyancy which is of gravitational origin and which would diminish the force of attraction.

this force diminishes with some power of the distance R from the sun. This is expressed by the equation

$$F = \frac{K}{R^n} \tag{2}$$

where the value of n is left open. Let us now examine the motion of one of the two planets. If we assume, for simplicity, that it follows a circular orbit (which is not far from the truth in most cases), we can express the centripetal force in terms of its mass and centripetal acceleration:

$$F_1 = m_1\omega_1^2 R_1$$

Since $\omega = 2\pi/T$, we can write

$$F_1 = m_1 \frac{4\pi^2}{T_1^2} R_1 \tag{3}$$

According to our assumption, the gravitational centripetal force can be expressed by equation (2). Substituting (2) in (3), we get

$$\frac{K_1}{R_1^n} = m_1 \frac{4\pi^2}{T_1^2} R_1 \tag{4}$$

or

$$\underline{T_1^2 = \left(\frac{4\pi^2 m_1}{K_1}\right) R_1^{(n+1)}} \tag{4a}$$

Equations (1) and (4*a*) are two ways of expressing the period of the planet 1 as a function of its distance R_1 from the sun times a factor which does not depend on this distance. Since both equations hold for any value of R_1, we can set $R_1 = 1$. We see then that the parentheses in both equations are equal to each other. If we now compare equations (1) and (4*a*), we see that, in general, $R_1^3 = R_1^{(n+1)}$

so that $n + 1 = 3$ and hence $n = 2$

Thus, we can write equation (2) as follows:

$$F = \frac{K}{R^2} \tag{2a}$$

This is the famous **inverse square law.**

Newton was also able to infer the value of K from Kepler's laws through considerations into which we cannot go in this text. The final form of the law of gravitation which he thus established is

$$\boxed{F = G\frac{m_1 m_2}{R^2}} \tag{13.1}$$

The force is proportional to each of the two interacting masses. The universal constant G, the so-called **gravitational constant** (not to be confused with g, the gravitational acceleration, which is not a universal constant), has the physical meaning of the force which is exerted by a unit point mass upon another unit point mass at unit distance. Its value depends on the choice of units of time, mass, and distance. In the cgs system its value is

$$\boxed{G = 6.670 \times 10^{-8} \text{ dyne-cm}^2/\text{gm}^2}$$

G is assumed to have the same value everywhere in the universe and to be independent of matter, fluid or solid, in which the two interacting masses are embedded.

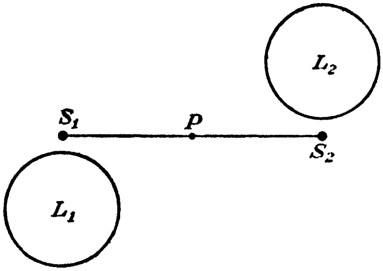


FIG. 13.3. Schematic diagram of Cavendish's experiment.

The value of G must be determined experimentally. The first determination was carried out by Cavendish in 1798 by means of the apparatus shown in top view in Fig. 13.3. A thin, light bar (S_1S_2) carries two small heavy metal spheres S_1 and S_2 at its ends. It is suspended by means of a fine fiber (which is to be visualized perpendicular to the plane of the paper) attached to the mid-point P. The small spheres S_1 and S_2 are attracted by the large lead spheres L_1 and L_2 according to Newton's law of gravitation. As a result, the suspended bar experiences a torque. This torque is measured by twisting the fiber so as to counteract the tendency of the spheres S_1 and S_2 to move toward L_1 and L_2. The torque due to twist, which is adjusted to equal the opposing gravitational torque, is found by measuring the angle through which the fiber must be twisted. In a preliminary experiment the fiber is calibrated by plotting the torque produced against the angle of twist.

In this experiment we know m_1, m_2, R, and F (from the torque), so that G can be found from equation (13.1).

But what did we mean by R—the distance between the fronts, the backs, or the centers of the spheres? R is the distance between the centers of the spheres. It is not at all self-evident that it should be so. When Newton conceived of his law of gravitation and performed his first computation relating to the motion of the moon about the earth, he assumed the masses of the two bodies to be concentrated at their centers and used the distance between the two centers as R in equation

(13.1). But in spite of a good agreement between the observations and the result of his computation, he felt that the agreement might be a coincidence and that his assumption might not be justified. It is an admirable example of the attitude of a conscientious scientist that Newton withheld the publication of this great discovery of his law for 20 years, until he was able to prove mathematically (after inventing the integral calculus) that it was indeed correct to treat the attraction between two spheres as if their entire mass resided at their centers.

From his law of gravitation and his three laws of motion, Newton was able to deduce all of Kepler's laws and to show in general that the path of a body influenced in accordance with the law of gravitation by another body must be a conic section. The different cases of conic sections are illustrated in Fig. 13.4.

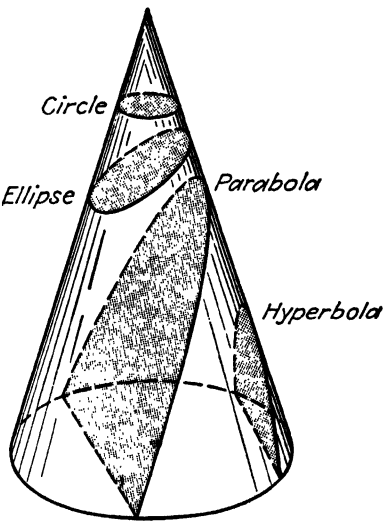


Fig. 13.4. Conic sections.

1. The **circle** is obtained by cutting at right angles to the axis of symmetry of the cone.
2. An **elliptic** section is obtained by severing the upper part of the cone, cutting at an arbitrary angle.
3. The **parabola** can be considered as a special case of the ellipse, the cutting plane being parallel to a straight line drawn on the cone surface through the apex of the cone. Its two branches tend to become parallel to each other at infinity.
4. A **hyperbola** is obtained by making the cutting plane more nearly parallel to the axis of the cone than in the case of the parabolic section. The two branches of the hyperbola approach two diverging lines at infinity (asymptotes).

Planets and their satellites, as well as double stars, revolve in circular or elliptic orbits. Orbits of many comets are nearly parabolic, so that these visitors of our solar system may make only a single appearance in our skies. One celestial body passing another will be deviated from its course by gravitational forces. Its path will be hyperbolic.

The above-mentioned cases are not only of astronomical importance. In the study of atomic structure we shall consider tiny electric particles (electrons) revolving in circular or elliptic orbits about a central body much as planets in our solar system revolve about the sun. And in experiments exploring the atomic nucleus we shall consider the deviation of α particles along hyperbolic orbits according to laws analogous to Newton's law of gravitation.

In addition to celestial phenomena, Newton also explained an important terrestrial phenomenon by his law of gravitation: the tides. Owing to the fact that the force of gravity diminishes in inverse proportion to the square of the distance, the face of the earth facing the moon is more strongly attracted toward it than the opposite side. This causes tides by deforming the water envelope surrounding our planet.

The Gravitational Field Intensity

Let us now return to the situation considered in the second section of this chapter. A mass released at rest at point A near a mass M will accelerate toward the center of M driven by a force F, which increases in inverse proportion to the square of the distance of the moving mass from the center of M. Wherever we place a test mass m, it will be acted on by a force directed toward the center of M. This leads us to assign to every point of space in the vicinity of M a vector, namely, the force exerted upon mass m. We have thus obtained a **force field,** a vector field, quite analogous to the velocity field in the case in which M would represent a sink in an extended fluid.

In order to be able to designate the strength as well as direction of the force field at every point, we shall define the concept of **field intensity** (or field strength) as follows: **the local field intensity is the force exerted upon a unit mass (1 gm) at the given point.**

Thus, if we are testing the field by measuring the force F exerted upon a mass m at a point, we find the field intensity I by forming the ratio

$$\boxed{I = \frac{F}{m}} \qquad (13.2a)$$

or at every point the force exerted upon a mass is given by

$$\boxed{F = Im} \qquad (13.2b)$$

In a gravitational field the force F has the meaning of the local weight W of the body. If we rewrite equation (13.2a) using this familiar symbol, we get

$$I = \frac{F}{m} = \frac{W}{m} = g$$

We recognize that $I = g$, that is, the gravitational field strength at any point is equal to the acceleration of gravity.

A vector field is difficult to work with. It would be much more convenient to associate a scalar quantity with the vector field such that the local value of the vector could be obtained from the scalar field when desired. In many cases such a scheme proves possible. It succeeds, for instance, in the case of certain types of flow, in the case of the gravitational-force field, and with certain types of electric field.

The Gravitational Potential

The association between a scalar and a vector field is suggested by our considerations of work and potential energy of Chap. 9. The potential energy of a body at a given point is measured with respect to a fixed level by the amount of work required to move the body from the fixed level to the point in question. No work is required in the gravitational field of the earth to move a weight along a horizontal surface. We called such surfaces **equipotentials,** or surfaces of constant potential energy. We also saw that the work done against gravity is independent of the shape of the path, being equal to force times vertical displacement (see Fig. 9.5).

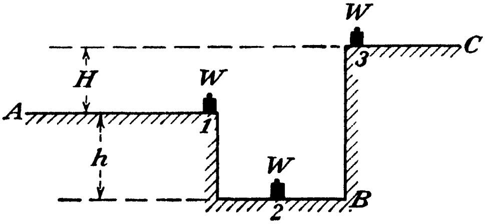


FIG. 13.5.

Let us consider, for example, Fig. 13.5. We can choose any of the three levels A, B, or C as our reference level to measure the potential energy of W. If we choose A, the potential energy of W in position 3 is positive; we have to do work to move W there; we can get the energy back when W falls from level C to level A. However, the potential of W in position 2 is *negative* with respect to A. If we choose the highest level C as our reference level, the potential energy of W will be negative at A as well as at B with respect to it. This change in value and sign of potential energy with changes in reference level has no profound significance; it is the consequence of the relative nature of the concept of potential energy.

Which would be the most convenient reference level for the measurement of gravitational potential energy? It is generally considered most convenient to use the level where the force vanishes, *i.e.*, the infinite height, as reference level. Thus, the potential energy at all points of finite height is negative, while the potential at the infinitely distant level is zero by convention.

Just as we used a unit mass of 1 gm to define the intensity vector in our field, we shall use it to assign a scalar quantity to every point of our gravitational field. We shall **define the *potential* at a point as the (negative) potential energy of a unit mass with respect to the infinitely distant level.**

At infinity our unit mass is said to have zero potential energy; as it approaches the source of the gravitational field, it loses potential energy and gains an equivalent amount of kinetic energy. Hence, we can also say: *The potential at a point is equal to the energy the unit mass would acquire by falling from infinity to that point*, or the potential is equal to the amount of work we have to do in bringing the unit mass back to infinity from the point in question.

In the following consideration, we shall ignore the sign of the potential, since it is a source of confusion for beginners.

Equipotentials and Lines of Force. Let us now assume that we have a mass M (Fig. 13.6) the gravitational potential field of which we are exploring. To simplify the problem, let us limit our consideration to one plane only.

We can determine (at least in thought) points of equal potential by releasing our unit mass from infinity and stopping it as soon as its kinetic energy has reached a certain prescribed value V_1. Thus, we shall obtain a series of points P_1, P_2, . . . , at which the unit mass arrives with equal kinetic energy from infinity. These will be equipotential points, and a line joining these points of equal potential will be an **equipotential line** (in space we shall obtain **equipotential surfaces**).

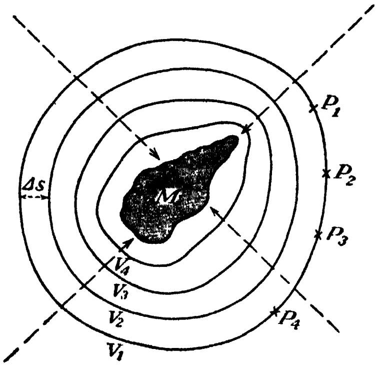


Fig. 13.6.

By varying the value of the kinetic energy at which we stop our particle, we shall get equipotential lines corresponding to different potential levels V_1, V_2, V_3, ...

Obviously, *no work is required to move a mass along an equipotential line*, since such changes of position do not involve a change of potential energy. This means, however, that the *gravitational force component along the direction of an equipotential line must be zero;* otherwise, work would have been done against this force component in moving a test mass along the line. In other words, the gravitational **force must be perpendicular to the equipotential lines at any point.**

Thus, we obtain the important result: Having established a series of equipotential lines (or surfaces), *we can draw the lines of force by drawing lines piercing the equipotentials at right angles.* The arrow of the force vector points in the direction of increasing negative potential.

Potential is a scalar (just like work). We have now established a way of obtaining the *direction* of the force from the scalar potential field (a field in which a potential value is assigned to every point in space).

Force and Potential. How can we now derive the *magnitude* of the force vector from the potential field? This is quite simple. Let us imagine we are moving a mass m across the small distance Δs in Fig. 13.6 (Δs being perpendicular to the equipotentials and being so small that the gravitational force may be regarded as practically constant in that interval). The force against which work is done in lifting m through the distance Δs is given by equation (13.2*b*): $F = Im$.

$$\text{Work done} = F\,\Delta s = Im\,\Delta s \tag{1}$$

The work done is stored in the form of potential energy. The change of the potential energy *of a* **unit** *mass* in transfer from equipotential line V_2 to V_1 is $\Delta V = V_1 - V_2$. And for the mass of m units, it is m times as large. Thus, the change of the potential energy of m is given by

$$E = m\,\Delta V \tag{2}$$

Equations (1) and (2) are two different ways of expressing the same energy changes. Hence, we can equate the right sides of these equations:

$$Im\,\Delta s = m\,\Delta V$$

or†

$$\boxed{I = \frac{\Delta V}{\Delta s}} \tag{13.3}$$

This is a fundamental equation. $\Delta V/\Delta s$ is called the **potential gradient.** It represents the change of potential per unit distance along a line of force (*i.e.*, at right angles to the equipotentials). Equation (13.3) states: **The magnitude of the field intensity I is equal to the potential gradient.**

Thus we can derive the magnitude and direction of the field intensity vector at any point of our field from the distribution of the scalar potential.

The concept of potential is of fundamental importance. It will be encountered in nearly all branches of physics.

---

† This equation is usually written with a negative sign to indicate that the force points in the direction of increasing negative potential energy.

QUESTIONS AND PROBLEMS

1. How do you obtain the field intensity at a point if the potential distribution in space is given?

2. What is a potential gradient?

3. How does the gravitational potential gradient of the earth change as you approach the earth?

★4. Two students somewhere in interstellar space are falling directly toward each other owing to their mutual gravitational attraction. The mass of Mr. A is 200 lb, and the mass of Miss B is 100 lb. Both students are equipped to measure their relative distance at any moment and have stop watches to measure time. Mr. A determines his acceleration a relative to Miss B and concludes that the force with which Miss B attracts him is $F_A = m_A a = 200a$.

Miss B finds the same acceleration relative to Mr. A, that is, a, and concludes that she is being attracted toward Mr. A according to Newton's law of gravitation with the force $F_B = m_B a = 100a$. According to this, $F_A = 2F_B$, but according to Newton's third law, both forces should be equal.

What mistake did the students make?

★5. Why does a body weigh less at the equator than at the North Pole?

★6. If the sun were completely dark, would there be at least a theoretical possibility of detecting the fact that the earth revolves about the sun?

7. Suppose you were suddenly transferred to a planet half as massive as the earth and one-fourth as large in diameter but otherwise having the same surface features. What physiological changes would take place in your body while you stand on the planet's surface?

8. Water issues from a point source and flows in all directions in a radial-flow pattern. If the local velocity of flow 3 ft from the source is 10 ft/sec, how large is the velocity of flow 9 ft from the source?

9. What is the force of gravitational attraction between a well-fed husband of 250 lb and his well-nourished wife of 200 lb? Both can be considered as spherical to a near approximation! Assume the distance between their centers to be 1 m. (1 lb mass = 0.45 kg mass.)

10. Assume in the Cavendish experiment (Fig. 13.3) that the mass of each large lead sphere L is 160 kg and of each small metal sphere 750 gm. Assume the distance between the centers of S_1 and S_2 to be 200 cm and between the centers of S_1 and L_1 21 cm. (These are the approximate dimensions of Cavendish's original apparatus.) How large is the torque exerted upon the arm S_1S_2?

11. In what way did the Cavendish experiment enable us to weigh the earth? How would you weigh the earth?

12. Does the earth have weight? Is this a meaningful question?

13. Can you determine the mass of the earth from the knowledge that 1 gm weighs 980 dynes and that the radius of the earth is 6,370 km?

14. Find the acceleration of gravity due to the earth in the vicinity of the moon, which is approximately 60 earth's radii away from the center of the earth.

15. Find the mass of the sun. Assume the orbit of the earth to be circular of radius $R = 1.5 \times 10^{13}$ cm. (Do you need the mass of the earth if it is negligible compared with that of the sun?)

16. How large is the gravitational field intensity at a distance from the earth's surface at which the acceleration of gravity is 75.3 cm/sec^2?

17. How large is the gravitational potential gradient at a point where a force of 5,000 dynes is exerted upon a mass of 10 gm?

18. The potentials corresponding to two closely adjacent equipotential surfaces are 25 and 30 ergs/gm. The perpendicular distance between the two surfaces in a certain region is 0.2 cm.

a. How large is the potential gradient?

b. How great a force is exerted upon a mass of 20 gm?

c. How much work is done in moving the mass of 20 gm from one equipotential surface to the other?

d. How large is the local field strength?

CHAPTER 14

FIELDS OF FORCE: 2. ELECTROSTATIC AND MAGNETOSTATIC FIELDS

a. In what way are the electric and magnetic forces similar to gravity?

b. In what way do they differ from it?

c. How can you magnetize an iron bar?

d. What is a magnet?

e. How can you demagnetize a magnet?

f. What are the poles of a magnet?

g. What makes you think the two poles are different?

h. How could you sever the north pole from the south pole?

i. What is the role of friction in electrifying a body?

j. Can you electrify glass by rubbing it against glass?

k. Can you electrify a gas?

l. What featutres of electricity suggested the concept of an electrical fluid?

m. Could the two-fluid hypothesis explain all the facts for which the one-fluid hypothesis accounts well? Is the converse true?

n. What observations led to the law of conservation of charge?

o. Can you charge one half of (1) a brass bar, (2) a glass rod leaving the other half uncharged?

p. What happens to the electrostatic attraction between two opposite charges if you transfer them from a vacuum into oil?

q. Could the repulsion between two like magnetic poles be increased by transferring them from air into an appropriate medium?

r. How do we define the electrostatic unit of charge?

s. Can you measure electrical charges or magnetic-pole strength by purely mechanical methods?

t. Is potential a mechanical or purely electrical concept?

u. What is an electric field?

v. What are lines of force?

w. On what else, besides the potential difference, does the work depend which you do in carrying a charge q from one point in a potential field to another?

x. What relationship exists between the field intensity and the variation of potential in space?

Magnetization

Although every day we see objects drop to the floor owing to the attraction by the earth, the similar attraction of a nail by a magnet or the lifting of little pieces of paper by an "electrified" comb across the intervening air gap still arouses our curiosity. The gravitational attraction is no less mysterious than these phenomena, but we simply get used to it in our early childhood and pay no more attention to it than to the air about us.

Whereas we do not have to treat a body in any special way to impart a weight to it, it is necessary to subject to a special treatment a piece of iron which is to be magnetized or a piece of glass which is to be electrified.

A piece of soft iron does not, in general, noticeably attract another piece of iron (except for a minute gravitational force which cannot be noticed without extreme experimental measures). But when we orient an iron bar approximately in the north-south direction and hit it end on with a hammer, it acquires the ability to attract small iron objects such as nails through a space gap. We say that the iron bar has become a **magnet.** By stroking another iron bar with our self-made magnet, we can "magnetize" the second iron bar, *i.e.*, convert it into a magnet.

Only a few materials, the so-called **magnetic substances**—*iron, cobalt, nickel* —display pronounced magnetic properties. They are strongly attracted by magnets and can be magnitized. Dropping or heating a magnet will erase its *active* magnetic property, *i.e.*, it will not attract soft iron any more but will still be attracted by a magnet.

Poles. If we dip a magnet into a bath of iron filings or small nails and take it out, a beard of small iron particles will adhere to it (Fig. 14.1). The beard will be longest near the ends of the bar and almost nonexistent near the center. We call the points at which the magnetic attraction seems to be concentrated the *poles* of the magnet.

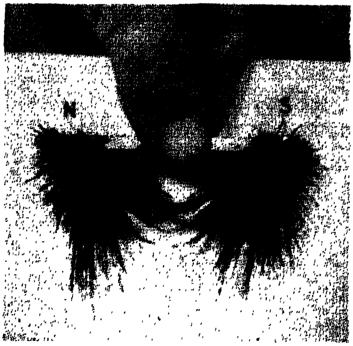


FIG. 14.1. Demonstration of the poles of a bar magnet through a "beard" of iron filings. (*From Weber, White, and Manning, "College Technical Physics,"* McGraw-Hill.)

If we now suspend one of our bar magnets by a fine string so that it hangs horizontally, it will orient itself approximately in the north-south direction. Evidently, the two poles are not equivalent; we call them the north-seeking and south-seeking poles, or north and south poles for short. Fine bar magnets (magnetic needles) pivoted on a pin point are used for purposes of geographic orientation (compass, Fig. 14.2).

All bar magnets behave as the compass needle showing two poles, **a north and a south pole.** *If we try to break off either of the poles, we don't succeed; we always get a shorter bar magnet with a north and a south pole.* **All magnets are dipoles** (have two poles).

If we now approach a pole of a magnet to a pole of a suspended magnet (such as the compass needle of Fig. 14.2), we notice that the **like poles**

(N and N, S and S) **repel each other, whereas unlike poles** (N and S) **attract each other.**

The geographic orientation of our suspended needle suggests that its north-seeking pole is attracted to the North Pole of the earth and its south-seeking pole to the South Pole. Thus, the earth behaves as if it were a magnet with a south-seeking pole at its geographic North Pole and a north-seeking pole at its South Pole.

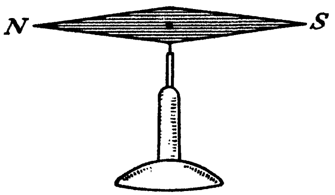


Fig. 14.2. A magnetic compass needle.

An unmagnetized piece of iron is always attracted by either pole and never repelled by a magnet.

Frictional Electricity

If we rub a celluloid ruler, for instance, with cloth, it acquires a property which is vaguely analogous to magnetism. It can attract small objects such as bits of paper, bread crumbs, iron filings, and brass filings. But there is an important difference between this process and magnetic attraction. The celluloid ruler exhibits its attractive properties with respect to any kind of matter, whereas pronounced magnetic effects are restricted to the so-called magnetic substances: iron, cobalt, and nickel. And there is no specific interaction between electrified bodies and magnetic poles.

When dipped into shreds of paper, the ruler shows no "poles." Its "electrification" is fairly uniform over its surface if we rubbed it uniformly. By rubbing only half of it, only half of the ruler will be electrified.

We can electrify not only celluloid but all kinds of matter without exception by bringing two dissimilar bodies into contact. Friction is not essential but advantageous. It serves merely to increase the area of contact between the two bodies, enabling microscopic projections of one to brush past the tiny crevices of the other. A glass rod rubbed with silk or sealing wax rubbed with fur behaves like the celluloid ruler. In fact, we find that the electrification is not confined to the glass and the sealing wax but that the silk used to rub the glass and the piece of fur which rubbed the wax also attract tiny particles.

Early Hypotheses on the Nature of Electricity

The minds of the early investigators of electricity were challenged by two basic questions: (1) In what way is a body changed when it is electrified? (2) How is the electrical force transmitted through space?

The action of a force without some "medium" of transmission, like a string or a rod, seemed inexplicable. The earliest hypotheses on the mechanism of such force transmission assumed that the air was acting as a medium. However, the invention of the air pump made it possible to prove that electric and magnetic forces, as well as gravity, were transmitted about as well through a vacuum as through air. The assumption of forces acting through empty space without a mediating material medium ("action at a distance") seemed unacceptable to natural philosophers. Their psychological need for a visualizable mechanism of force transmission was so strong that they were willing to assume the existence of a hypothetical "ether" (undetectable through our senses) permeating all matter and present even in the best vacuum, serving as a medium for the transmission of electric, magnetic, and gravitational forces. Descartes even devised an elaborate vortex hypothesis in the attempt to "explain" by a mechanical scheme the revolution of celestial bodies and their attraction toward the sun. He assumed a "celestial fluid" in a state of vortex motion carrying the planets around in circles. The attraction toward the center of rotation could be "explained" by the familiar experience that tea leaves tend to collect at the center of the vortex which is produced by stirring the tea with a spoon. Ingenious as Descartes' hypothesis was, it was eventually given up as fruitless and untenable along with other attempts to provide a mechanical explanation for the transmission of force through empty space.

To the question "in what way is a body changed by electrification?" one also sought to provide a mechanical answer or at least a description in terms of familiar phenomena. A clue seemed to be provided by the fact that electrified objects do not remain electrified forever. Eventually they seem to lose their electrification somehow. This gave some early investigators the idea that the "electrical virtue" was due to an electric fluid present in electrified bodies, which slowly evaporated from the electrified objects. Others (like Gilbert), on the other hand, assumed that the "electric fluid" was normally contained in all bodies in a spongelike fashion and that frictional heating made the fluid emanate from the body and surround it, forming an atmosphere of "electrical effluvium" which was supposed to mediate electrical forces "as air mediates gravitation."

Gilbert's successors imagined the electric fluid atmosphere to spin in a vortex about an electrified or "charged" body. The attraction of small objects was explained the same way Descartes "explained" gravitation by his vortex theory. A death blow was, however, dealt the vortex hypothesis when Du Fay discovered that electrified or "charged" objects could exhibit repulsion as well as attraction whereas the vortex hypothesis was capable of explaining attraction only.

The Two-fluid Hypothesis. If we suspend our "charged" glass rod as we did the bar magnet, it shows no geographic orientation. Another glass rod electrified in the same fashion repels it. And similarly, repulsion is found between two electrified celluloid rulers and between two electrified wax bars. It is equally easy to show that charged bodies can exhibit mutual attraction as well as mutual repulsion. A simple experiment teaches us that a sealing-wax bar electrified by rubbing with cat's fur is attracted by a glass bar electrified by rubbing with silk. We see thus that charged bodies may repel some charged bodies and attract others. One must, therefore, make some complicating assumptions in order to reconcile the fluid hypothesis of electricity with the additional observed facts. Du Fay suggested that there might be two kinds of electrical fluid, which he called *vitreous* and *resinous* electricity. The mutual repulsion of equally electrified bodies could then be expressed in terms of this hypothesis as follows:

Like fluids (or charges) repel each other.

All bodies which attract each other are then considered as charged with unlike fluids (or "charges"), and we can say:

Unlike charges attract each other.

We shall adopt the following definitions: (1) **All charges which repel each other are like charges.** (2) **All charges which attract each other are unlike charges.** This gives us a criterion to distinguish between charges of the two kinds.

The above statements are our *definitions* of what we mean by like and unlike charges. A definition cannot be false and, hence, should need no experimental verification. However, in this case we do have to perform experiments before we decide to adopt the above definitions for the following reasons:

In the above definitions we use the word "like" which, in common language, is associated with the following properties: (1) A is like A; (2) if A is like B, B is like A; (3) if A is like B and B is like C, A is like C.

Now if these rules do not apply to the meaning of the word "like" as used in our definitions of like and unlike charges, we should select a different word instead of like to avoid confusion.

Only experiment can decide if the meaning of the word "like" as used in the above definitions is synonymous with its everyday meaning. Our criterion for likeness of charges is as follows: Charge A is like charge B (written $A \sim B$) if and only if A repels B.

We see immediately that we cannot test experimentally statement (1) $A \sim A$, since A cannot repel itself. We therefore define the statement $A \sim A$ to mean A is charged or uncharged.

We don't have to experiment to check statement (2) in its application to charged bodies if we know that Newtonian mechanics applies to electrostatic forces; for $A \sim B$ means A repels B and we know from Newton's third law that B then equally repels A.

Statement (3) is the only one which must be tested experimentally. We must show that, whenever charge A repels charge B (meaning $A \sim B$) while charge B repels charge C (meaning $B \sim C$), charge A will repel charge C (meaning $A \sim C$). It turns out that what we defined as like electrical charges do show this behavior; hence, we can accept the above definition without danger of confusion.

But how do we know that there are only two unlike kinds of charges? We can show that all electrified bodies can be grouped into two classes: (1) those which are repelled by electrified resin and whose charge is like that of resin according to our definition and (2) those which are repelled by electrified glass to which we refer as being charged with vitreous (glasslike) electricity. The assumption of only two kinds of electricity (resinous and vitreous) has led to no inconsistencies and, hence, has been retained.

The Law of Conservation of Charge. Whenever we rub body A against body B, both become electrified with unlike charges. Benjamin Franklin proposed to call the two kinds of electrical charges **positive** and **negative** instead of vitreous and resinous. He was led to this idea by the observation that the two kinds of charge neutralize each other's effects. For instance, if you rub a piece of glass with silk and separate them, each will attract a distant small body. But if you wrap the silk about the glass, no attraction will be noticed. Take the silk off the glass and the first effect reappears. Thus, he reasoned that the amounts of electricity on the bodies A and B add up to zero, just like -7 and $+7$. He thus discovered the **law of conservation of charge:**

The amounts of electricity separated by contact of two bodies are always equal and of opposite sign.

In other words, *you cannot generate some place a charge $+Q$ without generating simultaneously elsewhere the charge $-Q$.* This version justifies our calling this a law of conservation of charge because it implies that the electrical charge of the universe must remain constant.

The One-fluid Hypothesis. Franklin was not happy about the two-fluid hypothesis. He felt that a theory assuming only one electrical fluid, besides being simpler, would also quite naturally explain the law of conservation of charge. He assumed that all bodies contained a certain "normal" amount of the electrical fluid which he identified with vitreous or positive electricity. When two bodies come in contact, he argued, one of them (body A) will lose some electrical fluid, thus becoming negatively charged, while the other (body B) acquires a surplus of electrical fluid (equal to the loss of body A), which makes it electropositive.

The modern view can be considered as a two-fluid hypothesis with a concession to the one-fluid hypothesis. Evidence to be presented in later chapters shows that all substances contain positive and negative electricity in equal amounts. In some bodies positive and negative electricity (or "electric fluids") are both freely movable, whereas in metals only the negative charges move about freely. Thus, the case of metals agrees most closely with Franklin's picture, with the exception that Franklin was wrong in assuming the movable electric fluid to be the positive one.

Conductors and Insulators

If we try to verify our previous statement that all bodies can be charged by "friction," we shall fail in the case of metals as long as we hold the metal object in our hand. But as soon as we put on rubber gloves, we shall succeed. The contact with our hand seems to "discharge" the metals or rather to prevent accumulation of charges on them. We find that we can discharge a metal object without touching it directly by hand. We can, for instance, touch it through a metal wire or a wet string. We call objects which allow the charge to leak away **conductors** of electricity. A rubber band or a silk thread will not discharge the body. These are so-called **insulators or dielectrics.**

If two metal bodies carrying equal but opposite electrical charges are bridged by a conductor, both bodies become electrically neutral. If they are bridged by a dielectric, they remain charged. Actually the dividing line between conductors and insulators is not sharp. Even an insulator bridge leads eventually to a discharge of the bodies, but it may take a long time, whereas a metal bridge acts instantaneously. There is a gradual transition from insulators to conductors.

Coulomb's Law

We have now come to a point where no important discoveries can be made by further *qualitative* considerations. We raise the following *quantitative* question: How does the electrical force of interaction depend (1) on the magnitude of the charges, (2) on their distance, (3) on the intervening medium?

Experiments designed to answer question (2) were carried out by Coulomb by an apparatus resembling the torsion balance of Cavendish (Fig. 13.3). The principle of Coulomb's apparatus is illustrated in Fig. 14.3.

The force of attraction between two small spheres carrying the charges

Q_1 and Q_2 tends to twist the fine suspension string S so as to diminish the distance d between Q_2 and Q_1.

The knob K can be turned so as to restore the original distance d between Q_1 and Q_2 (*i.e.*, the distance between the spheres in the uncharged state). The angle through which the knob K has to be turned is a measure of the force of interaction between Q_1 and Q_2. Coulomb showed with his "torsion balance" that **the force is inversely proportional to the square of the distance** d (provided d is large as compared with the radius of the charged spheres).

But how shall we determine the relationship between the force of interaction and the magnitude of the charges Q_1 and Q_2? We have not even defined how the magnitude of a charge can be measured!

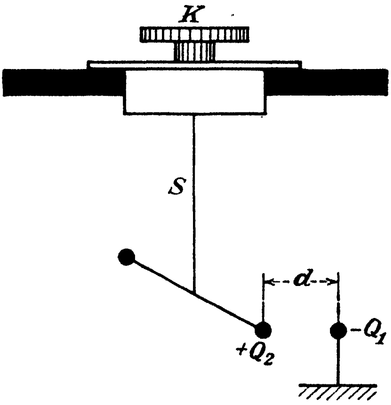


FIG. 14.3. Schematic diagram of Coulomb's torsion balance.

The only property of a charge known to us is its ability to exert a force upon another charge. Thus, we adopt the procedure of measuring the charge Q_1 on a body by determining the force F which Q_1 exerts upon a test charge Q_2. We define doubling of the force F exerted upon Q_2 to mean doubling of the charge Q_1 or, more generally, *we define the charge Q_1 on a body to be proportional to the force which it exerts upon the test charge.*

We know from experiments performed with the apparatus of Fig. 14.3 that $F \propto (1/d^2)$. From our definition of the magnitude of the charge Q_1, we know that $F \propto Q_1$ by definition. Hence, $F \propto (Q_1/d^2)$. But we can equally consider Q_1 as our test charge and vary Q_2, in which case $F \propto Q_2$, so that we can write $F \propto (Q_1Q_2/d^2)$. **The force is proportional to both charges, *i.e.*, to the product Q_1Q_2.**

Thus, for interaction *in a vacuum* we can write

$$F = C\frac{Q_1Q_2}{d^2}$$

What is the value of the proportionality constant C? This depends on the units we choose for Q. We can make C equal to 1, if we wish, as follows:

Define the unit of charge ($Q = 1$ esu†) as a point charge, which repels an identical charge 1 cm away in a vacuum with a force of 1 dyne.

---

† The abbreviation esu stands for electrostatic unit.

Then $F = 1 = C(1 \times 1/1^2)$ and $C = 1$. Thus, using this so-called *electrostatic unit of charge*, we can write **Coulomb's law for charges in a vacuum:**

$$\boxed{F = \frac{Q_1Q_2}{d^2}} \qquad (14.1a)$$

The practical unit of charge 1 Coulomb = 3 × 10^9 esu of charge.

Does the force change if the charges are, for instance, submerged in oil? Cavendish found that it does: *The force is largest in a vacuum; all material media reduce it.*

This state of affairs is expressed by multiplying the vacuum value of the force by a constant factor $1/K$. **K is called the dielectric constant. Its value depends on the medium.** Thus, we obtain the **generalized form of Coulomb's law for dielectric media:**

$$\boxed{F = \frac{1}{K}\frac{Q_1Q_2}{d^2}} \qquad (14.1b)$$

$K > 1$ *for all substances. It is very nearly equal to* 1 *for air and all gases. It is particularly large for water* ($K \approx 81$) (very much larger than for other common substances. This will be of significance in explaining certain properties of aqueous solutions on which the functions of living organisms depend.)

Coulomb's law holds strictly only for point charges, i.e, for charges which are spread over a volume whose linear dimensions are negligible as compared with the distance d between the charges.†

A similar quantitative law has been shown to hold for the interaction of magnetic poles:

Coulomb's law for magnetism: $\boxed{F = \frac{1}{\mu}\frac{p_1p_2}{d^2}}$ (14.2)

In this equation p_1 and p_2 are the strengths of the interacting poles measured in units defined similarly to the above-mentioned esu of charge.‡ The **permeability** μ is equal to 1 for vacuum. Media for which $\mu > 1$ are called **paramagnetic,** and in cases of extremely large values of μ (iron, cobalt, nickel) the medium is called **ferromagnetic.** Media for which $\mu < 1$ are called **diamagnetic.**

---

† For charges distributed over the surface or the volume of small spheres, the distance d in Coulomb's law is taken as the distance between the centers of the spheres.

‡ A unit magnetic pole is a pole which repels an identical pole a distance of 1 cm away in a vacuum with a force of 1 dyne.

Coulomb's laws are identical in form with Newton's law of gravitation, but the forces are of an entirely different order of magnitude. We have no difficulty demonstrating electric and magnetic attraction between electrified or magnetized bars. But the demonstration of gravitational attraction between such objects is exceedingly difficult. For example, two electrostatic units of charge 1 cm apart in a vacuum exert upon each other the force of 1 dyne. Two units of mass (1 gm) under the same conditions interact with a force of 6.7×10^{-8} dyne.

The establishment of Coulomb's law for electricity and magnetism is not merely of practical importance in measuring electric charges and magnetic poles. *Coulomb's law is of fundamental theoretical importance because it enables us to define and measure electric and magnetic quantities on a purely mechanical basis. It is the first necessary step in our aim of applying the laws of mechanics to phenomena of electricity and magnetism.*

The Leaf Electroscope

One might think of using the Coulomb torsion balance shown in Fig. 14.3 as an "electroscope," *i.e.*, a device for the detection of the presence of an electric charge on a body. For instance, if we were to touch a charged sphere Q_1 to the uncharged sphere Q_2, the charge of Q_1 would be distributed over both spheres and they would move apart owing to mutual repulsion, thus indicating that Q_1 was charged. But this apparatus is unnecessarily elaborate for such a purpose. The detection of a charge on a body is most commonly accomplished by means of leaf electroscopes of which Fig. 14.4 shows a common variety. A metallic plate P is connected through a metal stem S to two very light gold or aluminum leaves L. The stem passes through a hole in plug I made of an excellent insulator (for instance, amber). The plug is fixed to the cylindrical metal housing H, which is connected to the ground G by means of a wire. When the plate P is touched to a charged object, the charge is communicated to both leaves, which repel each other owing to their like charge and spread apart. The amount of divergence of the leaves depends on the charge of the electroscope. We shall make frequent use of this instrument in our subsequent discussion of electrostatics.

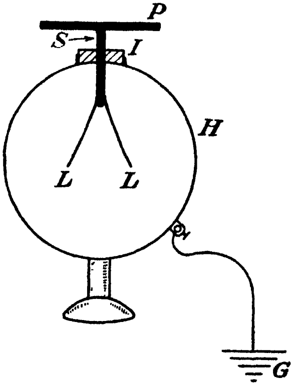


Fig. 14.4. A leaf electroscope.

The Electric Potential

We shall now take a second important step in the direction of creating a mechanical foundation for electrical concepts.

Suppose we charge a balloon of radius $R = 10$ cm made of conductive rubber. We connect it to an electroscope and note the deflection of the leaves. Now we deflate the balloon to a radius of 5 cm *without allowing any of its original charge Q to leak off.* We notice an *increase* in the electroscope reading. Although Q has remained constant, some electrical property of the balloon must have changed. Let us call this property for the time being the "degree of electrification." Although we have given this newly discovered property a name, we don't understand its meaning; *i.e.*, we have not correlated it with some familiar (preferably mechanical) concept. Let us attempt this correlation now:

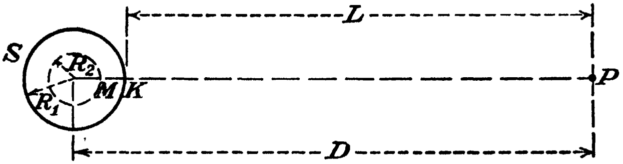


FIG. 14.5.

What we called the "degree of electrification" was increased by shrinking of the balloon. Has anything else increased at the same time? Figure 14.5 shows the balloon before and after its contraction (R_1 and R_2 being the original and the final radius, respectively). P is a point so distant from the balloon that we may refer to it as a point at infinity; *i.e.*, the radius of the balloon is negligibly small as compared with OP and the Coulomb force at P is imperceptibly small.

Now as the balloon shrinks, we see that the distance of the nearest point of its surface from P increases. This is our first correlation with the increasing "degree of electrification." Increase of the distance of the balloon surface from P implies an increase in the potential energy of a charge at the surface of the sphere with respect to the point P. This is easy to see. Suppose we have a point charge q at the original surface of our sphere ($R = 10$ cm) at the point K. This charge possesses potential energy. When it leaves the surface of the sphere, the charge Q remaining on the surface of the large sphere S repels the charge q as it moves toward P with a repulsive force which diminishes in inverse proportion to the square of the distance between the centers of the two spherical charges.

At P this force is practically zero and q has gained kinetic energy which is the equivalent of the potential energy it possessed before leaving the surface of the sphere S.

Now if we repeat this experiment after contraction of the sphere S, we see that the charge q will have to travel, not only the distance L as before, but also the distance $MK = R_1 - R_2$. It will have to move 5 cm farther and, *which is more important,* over a region which is closer to the center of the balloon and where the electric field is stronger than at any point of the previous path. More work will be done on the charge q now, and it will arrive at P with a higher kinetic energy. *This means that it possessed a higher potential energy on the surface of the sphere S after contraction than before contraction.*

Thus we have succeeded in correlating the *mechanical concept* of potential energy with our vague, preliminary *electrical concept* of "degree of electrification." For convenience we may choose $q = 1$ esu and **define as the electrical potential at a point the potential energy of a positive unit charge at that point relative to infinity.**

This means that the electric potential at a point in an electric field is the amount of work done on a unit positive charge by the electric field as the charge moves from the point to infinity. This amount of work can be shown, as a consequence of Coulomb's law, to be independent of the shape of the path along which the charge moves to infinity.

The potential at a point is unity if it requires 1 erg of work to move a unit positive charge from infinity to that point. **The electrostatic unit of potential (or of potential difference) is 1 erg/esu of charge. It is called 1 statvolt.** The potential difference between two points is 1 statvolt if we have to do 1 erg of work in moving 1 esu of charge (called 1 statcoulomb) from one point to the other.

The practical unit of potential difference is 1 volt = $\frac{1}{300}$ statvolt.

The volt is defined as follows: The potential difference between two points is 1 volt if we have to do 1 joule of work in moving the charge of 1 coulomb from one point to the other. [Since 1 joule = 10^7 ergs and 1 coulomb = 3×10^9 esu, we get from work = ΔVQ:

$$10^7 \text{ ergs} = (1 \text{ volt}) \times (3 \times 10^9 \text{ esu}); \; 1 \text{ volt} = \frac{10^7 \text{ ergs}}{3 \times 10^9 \text{ esu}} = \frac{1}{300} \frac{\text{erg}}{\text{esu}} \Big]$$

Since the potential difference ΔV represents the amount of work which must be done in moving a *unit* charge from one point to another, the amount of work done in moving a charge Q is given by

$$\boxed{\text{Work} = Q\,\Delta V}$$

We can now describe our initial observation in terms of the new concept of potential: *The potential of the balloon has been increased by diminishing its radius.*

We shall use from now on for the description of the electric state of bodies two quantities: the quantity of electricity (or charge) and the potential, as, in a similar fashion for the description of the thermal behavior of bodies, we shall use the quantity of heat (in calories) and the temperature (see Chaps. 15 and 17). Just as heat flows from points of higher temperature to those of lower temperature, positive electricity has a tendency to move from points of higher potential to those of lower potential. Negative charges move in the opposite sense.

Field Strength and Potential Gradient

We can now transfer to the field of electricity the results of our potential theory which we worked out for the gravitational field in Chap. 13. *M* in Fig. 13.6 may now represent a charged body, and the equipotentials may be lines of constant electric potential. Going through identical reasoning as in the derivation of equation (13.3), we obtain

$$\boxed{\mathcal{E} = \frac{\Delta V}{\Delta s}} \tag{14.3}$$

In this equation $\mathcal{E}$ is the **electrical field strength** and $\Delta V/\Delta s$ is the **electric potential gradient.**

The electric field strength $\mathcal{E}$ is defined similarly to the gravitational field intensity. **The electrical field intensity at a point is equal to the force exerted upon a unit positive charge at that point:**

$$\boxed{\mathcal{E} = \frac{F}{q}} \quad \text{or} \quad \boxed{F = \mathcal{E}q} \tag{14.4}$$

and similarly **we define the magnetic field intensity** *H* **as the force exerted upon a unit north pole:**

$$\boxed{H = \frac{F}{P}} \quad \text{or} \quad \boxed{F = HP} \tag{14.5}$$

The unit of electrical field intensity is 1 dyne/esu of charge.

The unit of magnetic field intensity is 1 dyne/unit pole. It is called 1 ***oersted.*** Since the field is proportional to the potential gradient ($\mathcal{E} = \Delta V/\Delta s$), we see that **a field-free region must be a region of constant potential,** *i.e.*, zero potential gradient.

Force Fields and Lines of Force

Having defined the vectors of magnetic and electric field intensity we may ask ourselves how we could obtain an idea about the distribution of the vectors in space, *i.e.*, about the appearance of the electric or magnetic vector field. In the study of flow we obtained an idea of the flow-field distribution by photographing a flowing suspension of particles, thus obtaining a streamline diagram (Fig. 13.1).

We can similarly obtain the distribution of a force field by a simple method. Suppose we arrange a positive and a negative charge in an oil bath in the places of the source and sink, respectively, of Fig. 13.1. A tiny positive test charge placed in the tank would tend to move (very slowly) through the oil, moving at every point very nearly in the direction of the field-intensity vector. The path will not be a straight one. Positive test charges released from different points around the stationary positive charge (source) will move along "lines of force" similar to the "lines of flow" we obtained with flowing water. At every point the direction of motion will be the same as the direction of the acting force. This direction can be obtained from a vector diagram representing the two forces exerted upon the test charge by the negative and positive charges which occupy the places of the source and the sink. The direction of the resultant varies from point to point (see example at the end of this chapter).

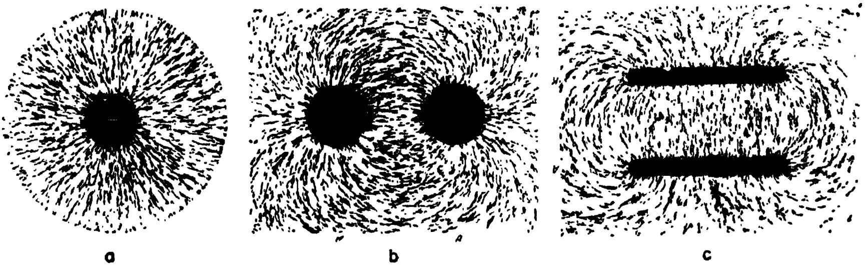


Fig. 14.6. Electric lines of force visualized with gypsum crystals which are oriented with their long axis parallel to the field. *a.* Example of an "inhomogeneous," *i.e.*, nonuniform, field. The lines of force converge toward the charge, indicating an increase of field strength as we approach the charge. *b.* Field of a dipole consisting of a positive and an equal negative charge. *c.* Field between two parallel oppositely charged plates. Field near center of plates is nearly "homogeneous" (uniform); *i.e.*, the intensity is nearly constant over a limited region of space in which the lines of force are parallel. The homogeneity decreases toward the edges of the plates. (*From R. W. Pohl, "Einführung in die Elektrizitätslehre," Springer.*)

A simple, practical method for the determination of electric "field lines" (or lines of force) consists in spraying small elongated crystals of epsom salt or gypsum on a glass plate which is placed in an electric field. They orient themselves (similarly to compass needles in a magnetic field) in the direction of the electric field and form chains which outline the electrical lines of force (Fig. 14.6).

Iron filings arrange themselves similarly in a magnetic field, depicting the magnetic-field distribution (Fig. 14.7).

Just as in the case of the fluid field the concentration (or "density") of the flow lines could be used as a measure of the local velocity of flow, the density of the lines of force measures the local field strength. Where the lines are most crowded, the field is most intense, and the number of force lines per square centimeter of cross section can be used as a measure of the field intensity (Fig. 14.8).

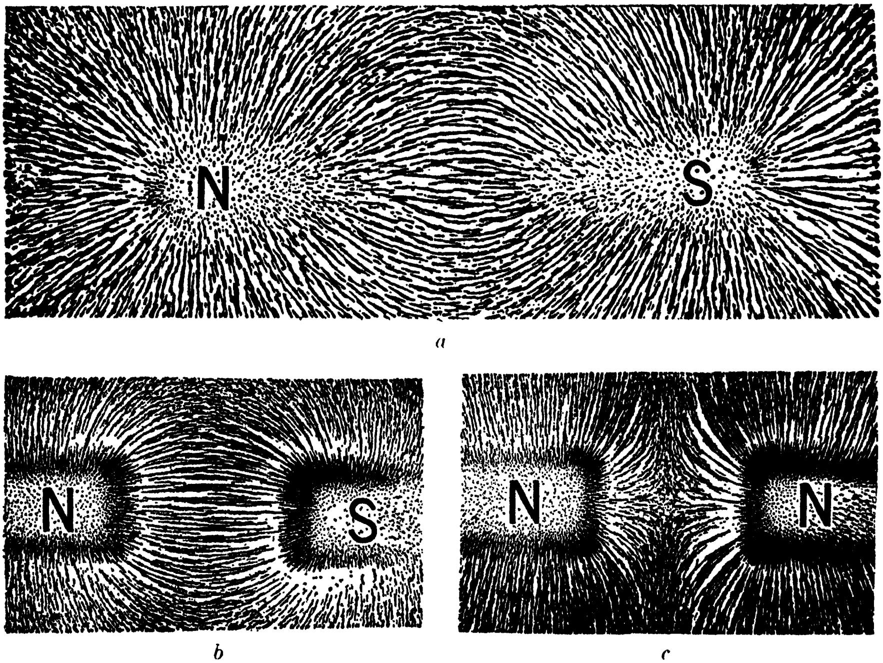


FIG. 14.7. Magnetic field lines visualized with iron filings. *a*. Field of a bar magnet. *b*. Field between two opposite poles of two bar magnets. *c*. Field between two like poles of two bar magnets. (*From Weber, White, and Manning, "College Technical Physics," McGraw-Hill.*)

Lines of force are merely lines which indicate at every point the direction of the field-intensity vector in a magnetic or electric field. Faraday, however, who was very active in searching for a mechanism of force transmission through space, ascribed to these lines of force actual physical properties. The behavior of charged bodies was described by Faraday as if they were attached to lines of force, having a tendency to contract individually while tending to repel adjacent lines. This is a useful mental picture which helps us to visualize and memorize certain facts but to which no physical reality is to be ascribed. It does not even enjoy the status of a hypothesis any longer.

Faraday's ideas on the mediation of electric and magnetic forces were formulated mathematically and developed further by Maxwell. He assumed space to be permeated with an elastic medium, the ether, which was supposed to transmit electric and magnetic forces (as well as light) and which was supposed to be in a state of elastic deformation in the region of a field, so that a certain content of potential energy of deformation could be associated with an electric or magnetic field. We shall see later that such mechanical models of the ether proved superfluous and, in fact, untenable.

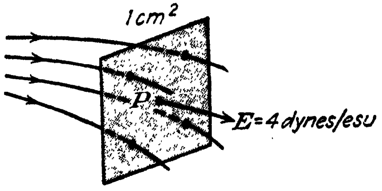


FIG. 14.8. Graphic representation of field intensity. The number of lines of force passing through the area of 1 cm² is numerically equal to the local field strength. (*From Freeman, "Modern Introductory Physics," McGraw-Hill.*)

APPENDIX 1

Some Medical Applications and Examples

1. Tiny particles suspended in air are electrically charged, mainly by "friction." This applies to liquid particles such as water droplets as well as to solid ones such as flour dust or coal dust, which may form charged clouds. Bacteria suspended in air may also acquire an electric charge. When the potential difference between oppositely charged bodies or clouds becomes high enough, the air between them may be rendered temporarily conductive (by a process to be discussed later) and the bodies will discharge each other through a spark. This phenomenon is observed on a large scale in thunderstorms and on a smaller scale when sparks jump from charged clouds of flour dust or coal dust to the ground, sometimes causing severe explosions of the combustible dust.

2. People walking over a carpet in winter may easily become charged and cause spark discharges to ground or to other people. By a similar process operating-room personnel using rubber-soled shoes are apt to acquire electrical charges. Hundreds of explosions of anesthetic gas mixtures (cyclopropane or ether mixtures with air) reportedly have taken place in operating rooms during operations, in many cases blowing up the lungs of the patient and injuring physicians and nurses. The cause of these explosions has been attributed to tiny sparks caused by electrically charged members of the operating crew. To avoid such occurrences one attempts to maintain all objects and persons near the anesthesia machine at the same potential. The personnel can, for instance, wear shoes of conductive rubber or with perforated soles so as to lose accumulating charges quickly to ground. In addition, a fairly high relative humidity is maintained in the operating room. If the air is humid enough, a surface layer of moisture usually condenses on objects. The surfaces of insulators are ren-

dered more conductive by the moisture layer and provide a conductive path for the escape of electric charges to ground.

3. Electrification of small particles can be also used to clean the air of air-borne bacteria and dust particles. The particles are charged by being passed over wires maintained at a high potential with respect to the ground. Then the air containing the charged particles is passed between two oppositely charged plates on which they are precipitated.

4. Electric-spark discharges through a patient have been widely used in therapy by physicians and enjoy great popularity among quacks. Benjamin Franklin was a pioneer in the field of electrotherapy but admitted his results to be inconclusive. This treatment is being used to cause involuntary muscle contractions in conditions such as sprains, contusions, or lumbago.

5. Magnetism also has some medical applications. For example, iron particles embedded in the eye constitute a frequent industrial accident. They are removed with a straight (electro)magnet equipped with a pointed cone at its end ("eye magnet").

APPENDIX II

Summary of Electric and Magnetic Units

1. **The Unit of Charge.** In the **electrostatic system** of units, the basis for defining a unit charge is Coulomb's law: $F = q_1q_2/Kr^2$. If two equal plus (or minus) charges placed 1 cm apart in a vacuum repel each other with a force of 1 dyne, each is said to have one electrostatic unit (esu) of charge (or 1 statcoulomb).

There are **two other systems** of units in use, the **electromagnetic** (emu) and the **practical.** The following relations exist between the three systems:

$$\begin{aligned}1\text{ emu of charge} &= 3 \times 10^{10}\text{ esu of charge}\\ &= 10\text{ coulombs (the coulomb being the practical unit of charge)}\end{aligned}$$

2. **The Magnetic Pole Strength.** We define the **unit pole** as the pole which, placed 1 cm away from an identical pole in a vacuum, repels it with a force of 1 dyne. In terms of this unit it is possible to make quantitative comparisons between the pole strengths of different magnets.

3. **The Unit of Potential Difference.** The potential difference (p.d.) between two points in an electrostatic field is the work done in moving a unit positive charge from one point to the other. **One esu of potential difference** (1 statvolt) exists between two points when 1 erg of work is done by the field in moving one positive esu of charge from the point of higher to the point of lower potential.

(In the **electromagnetic system,** the unit of potential difference is defined similarly in terms of the erg and of the emu of charge.)

In the **practical system of units,** the unit of potential difference is the volt. One volt is the potential difference that exists between two points if 1 joule (10^7 ergs) of work is done in moving a coulomb of charge from one of the two points to the other.

The three systems are related as follows:

$$\boxed{\begin{aligned} 1 \text{ esu of p.d.} &= 3 \times 10^{10} \text{ emu of p.d.} \\ &= 300 \text{ volts} \end{aligned}}$$

4. **Capacitance** (see Chap. 15). Capacitance (or capacity) is defined as the ratio

$$C = \frac{Q}{V} = \frac{\text{charge on a conductor}}{\text{potential of the conductor}}$$

The unit of capacitance is equal to the unit of charge divided by the unit of potential difference. In the electrostatic system, the unit is the centimeter. This is the capacitance of a body which is charged to the potential of 1 esu of potential difference by 1 esu of charge.

In the practical system, the name of the unit is the farad.

$$\boxed{1 \text{ farad} = 1 \text{ coulomb/volt} = 9 \times 10^{11} \text{ cm}}$$

5. **Electric Field Strength.** The electric field strength, or electric intensity, at a given point is defined as the force exerted on a unit positive charge placed at that point. **Hence, it is measured in the electrostatic system in units of dynes per esu of charge.**

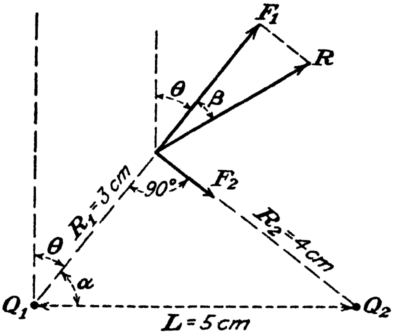
An alternative way of defining the unit of field strength is based on equation (14.3). *The unit field strength exists at a point where the potential gradient is unity, i.e.,* 1 *statvolt/cm.* It is easy to see that this unit is identical with 1 dyne/esu of charge if we remember that 1 statvolt = erg/esu and that 1 erg = 1 dyne-cm. We can write unit field strength $= 1\ \dfrac{\text{statvolt}}{\text{cm}} = \dfrac{\text{erg/esu}}{\text{cm}} = \dfrac{\text{dyne-}\cancel{\text{cm}}}{\text{esu } \cancel{\text{cm}}} = \dfrac{\text{dyne}}{\text{esu}}$. *The practical unit is* $1\ \dfrac{volt}{cm}$.

6. **Magnetic Field Strength.** The magnetic field strength or magnetic field intensity is defined for a given point as the force exerted at that point upon a unit magnetic north pole. The unit is 1 **oersted** (1 dyne per unit pole) if the force is measured in dynes and the pole strength in unit poles.

Example

Let us assume that we have two charges $Q_1 = +30$ esu and $Q_2 = -16$ esu placed $L = 5$ cm apart, as shown in the diagram. It is desired to determine the electric field in the space surrounding the two charges

Solution. One might think at first of an experimental solution of this problem. By placing a small charged object at various points near the charges and determining the magnitude and direction of the local force, one could map a picture of the lines of force roughly similar to Fig. 14.6*b*. This experimental method is, however, quite impractical. A simple theoretical method based on Coulomb's law yields the field-intensity vectors at any point with great precision and simplicity as follows:



The point P at which we shall determine the field strength has been chosen conveniently to simplify the numerical computation. Suppose a charge of $q = +30$ esu to be placed at point P. This test charge will be acted on by two forces: a repulsion from Q_1,

$$F_1 = \frac{Q_1 q}{R_1^2} = \frac{30 \times 30}{9} = \underline{100 \text{ dynes}}, \text{ and an attraction toward } Q_2,$$

$$F_2 = \frac{Q_2 q}{R_2^2} = \frac{16 \times 30}{16} = \underline{30 \text{ dynes}}.$$

We see that the resultant force R acting upon the charge q at the point P is directed neither away from Q_1 nor toward Q_2. Its magnitude and direction follow from the force triangle (which is in this case rectangular).

$$\underline{R} = \sqrt{F_1^2 + F_2^2} = \sqrt{(100)^2 + (30)^2} = 10\sqrt{109} = \underline{104.5 \text{ dynes}}$$

The field intensity at that point, being the force per unit charge, is

$$\underline{E} = \frac{R}{q} = \frac{104.5}{30} = \underline{3.48 \text{ dynes/esu}}$$

The direction of R (and hence also of the field strength E) is also found easily. Its angle with the vertical is $(\theta + \beta)$ as can be seen from the figure.

$$\theta = 90° - \alpha$$

$$\sin \alpha = \tfrac{4}{5}; \qquad \alpha = 53°10'; \qquad \theta = 90° - 53°10' = 36°50'$$

$$\tan \beta = \frac{F_2}{F_1} = \frac{30}{100}; \qquad \beta = 16°40'$$

hence,

$$\underline{(\theta + \beta) = 53°30'}$$

Using this procedure we can find the field vector $\vec{E}$ at any point. By determining $\vec{E}$

(or $\vec{R}$) for several consecutive points which are selected so that each point is removed from the preceding point a short distance along the line of the resultant, the student will get an understanding of why the lines of force are curved in this case.

QUESTIONS AND PROBLEMS

1. What evidence led to the assumption (*a*) of different kinds of electricity? (*b*) of two kinds of magnetic poles?

2. If you do twice as much frictional work in rubbing a glass bar with silk, will you get twice as high a charge on the glass?

3. Assume the following "one-fluid" hypothesis: Electrification of a body is due to excess or deficit of a compressible electric fluid which permeates the body. When the fluid is in excess, the body is positively charged; when there is a deficit in electric fluid, the body is negatively charged.

Can you propose experiments which would test whether or not this hypothesis is tenable? Will there be any upper or lower theoretical limit to the negative or positive charge which the body could acquire.

4. How do we know that the force between two point charges is proportional to their product? Is this knowledge based on experimental results?

5. How can you justify the statement that two lines of force cannot cross each other?

6. What properties did Faraday ascribe to the lines of force?

7. Where do the electric lines of force originate and where do they end?

8. Compute the force of interaction between the following bodies placed 1 cm apart in a vacuum.

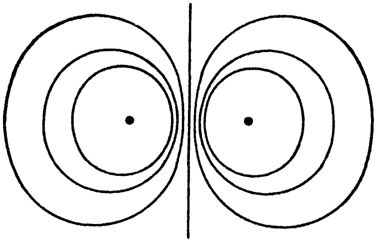
a. Two point masses 1 gm each.

b. Two positive point charges 1 esu of charge each.

c. Two magnetic unit north poles.

d. Two positive point charges 1 coulomb each.

9. The diagram shows the equipotential lines of a dipole (two equal and opposite electric charges or magnetic poles). Draw the corresponding lines of force on a transparent sheet of paper placed over this diagram.



10. A magnetic compass needle of 5-cm pole distance and pole strength of 15 unit poles per pole is placed in a magnetic field where the magnetic field strength does not vary from place to place. The value of the constant magnetic field intensity is 20 oersteds. The axis of the needle is perpendicular to the lines of force.

a. Make a sketch of the lines of force of this field and of the magnetic needle.

b. How large is the torque exerted upon the needle?

11. The magnetic field intensity at a distance x from a given magnetic pole is 20 oersteds. The strength of the given pole is 40 unit poles. Find the distance x.

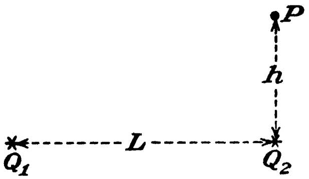
12. A charge of $q = -20$ esu is placed between a charge of $Q_1 = +100$ esu and $Q_2 = +40$ esu so that the three charges lie on a line.

a. At what point between Q_1 and Q_2 will the charge q remain in equilibrium?

b. Will this equilibrium be stable?

c. If we replace q by an equal but opposite charge, will it be in equilibrium at the same point and will its equilibrium be stable?

13. Determine the electric field intensity in magnitude and direction at the point P. $Q_1 = 100$ statcoulombs; $Q_2 = -100$ statcoulombs. $L = 10$ cm; $h = 5$ cm.

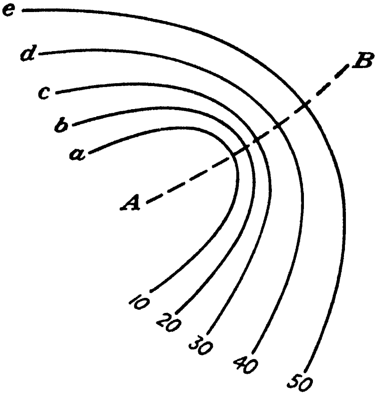


14. A charge of +25 esu is carried in an electric field from a point whose potential is 50 volts with respect to ground to a point maintained at 80 volts relative to ground.

a. How much work is done on the charge?

b. How much power is required to transport electric charges at the rate of 25 coulombs/sec from one of these points to the other?

15. *a, b, c, d* are sections of equipotential lines. *AB* is a section of a line of force. The segments of the line of force cut off by the equipotential lines are as follows: $ab = 10$ mm, $bc = 8$ mm, $cd = 6$ mm. The numbers labeling the lines *a, b, c, d* represent their potential in volts.



a. Find the average field strength along the line *AB* between adjacent equipotential lines.

b. How large a force is exerted upon a charge of 5×10^{-10} esu placed centrally between lines *b* and *c* along the line of force?

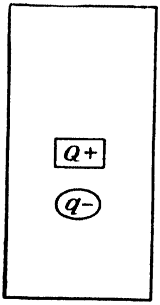
c. How much work is done in moving a charge of 10 esu from line *a* to line *d*?

16. Compute the acceleration of the following bodies placed in a homogeneous electric field (*i.e.*, a field whose lines of force are parallel and whose intensity does not vary from place to place) of the field intensity of 100 dynes/esu of charge.

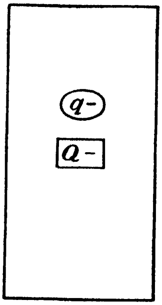
a. A mass of 1 gm charged with 1 statcoulomb.
b. A mass of 1 gm charged with 1 coulomb.
c. A mass of 10^{-24} gm charged with 5×10^{-10} statcoulomb.
d. A mass of 10^{-27} gm charged with 5×10^{-10} statcoulomb.

17. How large is the attraction between two point charges of +20 esu and −100 esu 5 cm apart in oil of the dielectric constant $K = 2.0$?

18. The equilibrium of a particle is said to be stable if the particle has a tendency to return to its original position after a slight deviation. In the following diagrams a small charge $-q$ is free to move up or down (but not sideways) near a large charge Q.



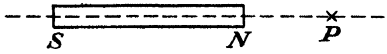
A



B

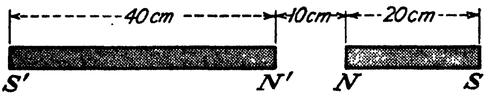
The particle is in equilibrium since the electrical upward force is equal to the weight of the particle. Which statement is true:
a. In Fig. *A*, $-q$ is in stable equilibrium.
b. In Fig. *A*, $-q$ is in unstable equilibrium.
c. In Fig. *B*, $-q$ is in stable equilibrium.
d. In Fig. *B*, $-q$ is in unstable equilibrium.
e. The equilibrium is unstable in both Figs. *A* and *B*.

19. A magnetic north pole P whose strength is 4 magnetic units lies on the axis produced by a bar magnet, 2 cm away from its north pole. The poles of the bar magnet



are 8 cm apart, and its magnetic moment is 80 magnetic units. How large is the force in dynes exerted on the isolated north pole?

20. Two magnets of the dimensions shown in the diagram interact with each other



in the position shown. Do they attract or repel each other?

a. Find the force of interaction if the magnetic pole strengths are as follows:

$$|S'| = |N'| = 30 \text{ magnetic units}$$
$$|N| = |S| = 20 \text{ magnetic units}$$

b. Find the force for the case that the smaller magnet is reversed, *i.e.*, S faces N'.

CHAPTER 15

FIELDS OF FORCE: 3. BEHAVIOR OF MATTER IN ELECTRIC AND MAGNETIC FIELDS

a. Does a stationary electric charge interact with a stationary magnetic pole?

b. Does an electric charge attract or repel nonelectrified matter? Explain why.

c. Can you charge two bodies simultaneously by induction?

d. Which of the two charges separated by induction is the larger, the one near the inducing charge or the distant one?

e. On what evidence do you base your answer to *d*?

f. Can you charge a single body by induction? Are you sure you are charging a single body?

g. Is the electrical attraction of a neutral insulator based on the same process as the attraction of a neutral conductor? On what evidence do you base your answer?

h. Using a positive inducing charge, what charge can you communicate to another body by induction?

i. In separating charges by induction (Fig. 15.1), would you first remove the rod R and then separate the bodies or would you proceed in the reversed order?

j. Can the magnetization of soft iron be explained by a fluid hypothesis of magnetism?

k. What is the maximum charge which a charge Q can induce in another body?

l. On what factors does the actual amount of charge induced in a body depend?

m. What did Faraday's ice-pail experiment prove?

n. What is a capacitor?

o. Does the potential of a charged surface depend on the electrical condition of its environment?

p. What is the meaning of the term "capacitance of a condenser"?

q. In what units is it measured?

r. Can the potential of a sphere be changed by changing its capacitance?

s. What effect do dielectric substances have upon capacitance?

t. What is meant by the term "dielectric polarization"?

u. What hypothesis is used to explain magnetization of iron?

v. Can any substance be magnetized?

w. In what ways do ferromagnetic, paramagnetic, and diamagnetic bodies differ from each other?

Electrostatic Induction

The following facts have been established experimentally: (1) Neutral masses exert gravitational forces upon each other; (2) like electric charges repel each other; (3) unlike charges attract each other; (4) like magnetic poles repel each other; (5) unlike magnetic poles attract each other; (6) stationary electric charges and magnetic poles do not interact; (7) an electrified body may attract neutral particles. Does the last statement mean that there is an interaction between electrical charges and gravitational masses? Before we draw such a conclusion, we should make sure that our "neutral" attracted particles are really neutral during the process of attraction. Experiments show that this is not quite the case. We arrange a neutral metal body B so that it can be separated into partial bodies B_1 and B_2 (Fig. 15.1). We approach a charged rod R to the body B while its parts B_1 and B_2 are in contact and then separate B_1 from B_2. We find with the aid of an electroscope the two "half bodies" to be charged with opposite kinds of electricity. When we remove the charged body R, the charges on B_1 and B_2 persist, but when we touch B_1 and B_2 in the absence of R, the body B becomes neutral again, indicating that the opposite charges on B_1 and B_2 were equal.

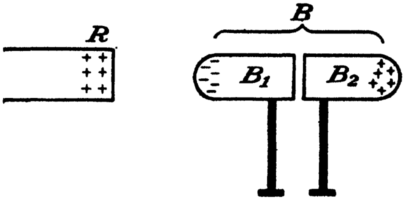


FIG. 15.1. Electrostatic induction in a body B which can be split into two "half bodies" B_1 and B_2.

Evidently the separation of charges took place in our conductor B owing to the proximity of the charged rod R, and it is quite easy to understand why. By "understand" we mean here to associate an imaginary mechanical process with the observed phenomenon through a hypothesis. Let us describe the process in terms of the two-fluid hypothesis: The positive rod R attracts the particles of the "negative fluid" in the body B, and they start moving through the conductor toward R (B_1 and B_2 being in contact). At the same time the particles of the "positive fluid" are repelled from R and collect at the opposite end of B. How long does this process of separation continue? Imagine a freely movable particle of the negative "electric fluid" in body B. At the beginning of the process it is attracted toward R. As more and more opposite charges are segregated at the ends of the "half bodies" B_1 and B_2, the force experienced by the charge in the interior of the body becomes smaller, since the negative cloud formed at the left end of B_1 repels the freely movable charge to the right while at the same time the positive cloud on the right end of B_2 attracts it toward the right. As this process of separation of charges continues, the positive and negative clouds are growing at the ends of B. A state is finally reached where the electric field which originally existed in the interior of the body B owing to the charge on R is completely neutralized by the opposite electric field which is set up by the positive and negative clouds of the electric fluid at the ends of B. As soon as this state is reached, all charged particles in the interior of the body B are in a *field-free space* and subject to no forces. The process of **electrostatic induction** (*i.e.*, separation of charges) is now completed; it is practically an "instantaneous" process.

As we have mentioned, in the case of metallic conductors only the negative charge is freely movable. The positive charge at the right end of B_2 is then not due to a migration of positive charges but, rather, to a depletion of negative charges which wander off toward B_1, leaving an equal amount of positive charge uncompensated.

If B were a hollow metal box, the opposite charges induced at its ends would neutralize the field created by the charged body R in the interior of the box. Points inside the box would then be *shielded* from disturbances by extraneous electric fields. Electrostatic shielding is widely used in connection with industrial and medical electronic equipment.

If B_2 were the terrestrial globe and B_1 a "grounded" body (*i.e.*, a body connected to ground by means of a wire), we should have charged body B_1 *negatively* by induction by breaking the contact with the earth. Unlike the case of **charging by contact,** where a charging body communicates part of its own charge to the body which it charges, the same body will **communicate a charge of opposite sign when charging a body by induction.** The inducing charge remains undiminished in this process.

It is now easy to see how an originally electrically neutral body is attracted by a charged one. In Fig. 15.1 B was neutral to begin with. When the charged bar R approaches B, the end proximal to R becomes negative and hence is attracted toward R; the distant end is positive and, hence, is repelled from R. Since the induced opposite charge is closer to R than the induced like charge, the force of attraction exceeds the repulsive force and the resultant force is always one of attraction.

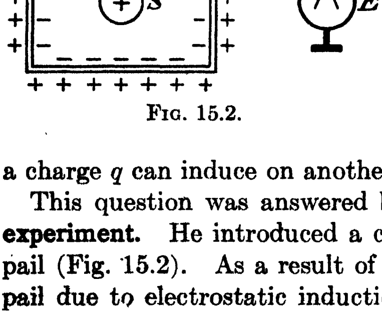


FIG. 15.2.

Magnitude of the Induced Charge. Let us ask now a quantitative question regarding electrostatic induction. What is the maximum charge which a charge q can induce on another body?

This question was answered by Faraday through his famous **ice-pail experiment.** He introduced a charged metal sphere S into a closed ice pail (Fig. 15.2). As a result of separation of charges in the wall of the pail due to electrostatic induction, the electroscope E showed a deflection. As a result of the field of the positive charge on the sphere S, negative charges are attracted toward the inner surface of the pail whereas the positive ones are repelled to its outer surface and charge the electroscope. The electroscope deflection does not vary as the sphere S is moved about inside the pail, but it disappears when the sphere S is taken out of the pail and reappears when it is reinserted.

When S is touched to the inside of the pail, the deflection of the electroscope E does not change, and it persists undiminished after S has been removed from the interior of the pail. An examination of S with an electroscope shows that S is *completely* discharged and the charge on the outside of the pail proves to be equal in sign and magnitude to the original charge of S. Thus, since the deflection of E did not change when S touched the pail and all of the charge of S was transferred to the pail, we conclude that the induced charge present on the outside surface of the pail before touching was equal to the inducing charge on the sphere S. **The total charge which a charged body can induce on its environment,** and hence the maximum charge it can induce on another body, **is equal to its own charge.**

This result can be deduced in the following simple special case (Fig. 15.3) from two verifiable assumptions, namely, that a uniform charge on a spherical surface (1) exerts forces in the region outside the sphere as if the charge were concentrated at the center of the sphere† and (2) exerts no forces in the region inside the spherical shell as if the charge were nonexistent.‡

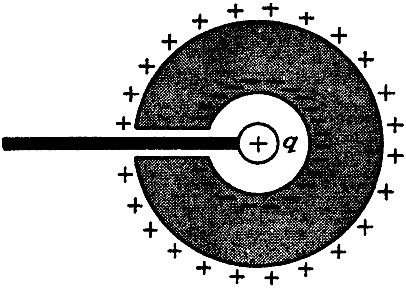


FIG. 15.3.

Imagine for simplicity a thick-walled spherical container made of conductive material: The charge $+q$ introduced in its interior will affect the positive and negative charges in the interior of the wall (we assume, for the sake of generality, both charges to be freely movable), attracting the negative ones to the inner surface of the container until q units of negative charge are accumulated on it. At that point the field inside the thick container wall becomes zero, since the negative charge on the inner surface acts like a charge $-q$ at the common center of the spheres, combining, as it were, with the inducing charge $+q$ at the center to give a zero effect. This process limits automatically the magnitude of the induced charge to the value of q. The like charge $+q$ has been at the same time repelled to the outside surface of the container from where it exerts no forces at interior points. If we touch the inducing sphere to the inner surface of the container, the inducing charge $+q$ will be neutralized by the negative charge induced on the inner surface and only the outer charge of $+q$ will remain. (These results are not restricted to spherical containers, as can be proved mathematically.)

---

† This theorem can be proved mathematically in a similar fashion to Newton's theorem stating that the gravitational force exerted by a spherical mass is such as would be exerted if the mass were concentrated at the center.

‡ This can be shown mathematically on the basis of Coulomb's law.

The same consideration shows also that **a charge cannot reside in the interior of a conductor.** It will immediately begin to attract opposite charges until it is neutralized, while an equal amount of like charges has been repelled to the outermost points, namely, the surface of the conductor. **The charge on a charged conductor resides on its surface. The static charge must be distributed on the surface in such a fashion as to make it an equipotential surface.** Suppose for a moment that a point at the surface is at a higher positive potential than its surface environment. The existence of the potential difference implies the existence of a potential gradient and hence of a force field which moves negative charges toward the positive "island of charge" until the plus charge has been neutralized and the potential gradient abolished. Similarly we can infer that the interior of the volume of a conductor must be at equal potential; *i.e.*, the internal field must be zero. The fact that the charge resides on the surface of a conductor can be demonstrated as follows: Figure 15.4 shows a sphere S mounted on an insulating stand. H_1 and H_2 are two hemispheres which fit the sphere S closely and are provided with insulating handles. When we charge S while H_1 and H_2 are removed and then place H_1 and H_2 over S as shown in the figure, the entire charge of S is transferred to H_1 and H_2. When H_1 and H_2 are removed from S, the electroscope shows S to be completely discharged and the two hemispheres to have acquired its charge. If, on the other hand, we charge the body while H_1 and H_2 are pressed against S as shown in the figure and then remove H_1 and H_2 from S, the two hemispheres prove charged and S has no charge.

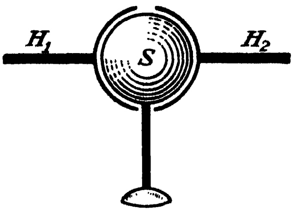


Fig. 15.4. Demonstration of the fact that static charges reside on the surface of a conductor.

Capacitors and Capacitance

The question now arises: Is the potential of a charged body affected by the charges which it induces in neighboring objects? A simple consideration shows that it must be. Figure 15.5 shows a charged sphere $+Q$. It has a certain potential; *i.e.*, we have to do a certain amount of work to bring a unit positive charge from infinity to the surface of Q. If we now place the long-handled metal dumbbell D near Q, a charge of $-q$ will be induced on the dumbbell sphere nearer Q. This obviously reduces the potential of $+Q$, since the positive unit test charge, which, in the absence of the dumbbell, advanced only against the repulsion of $+Q$ in approaching this charge, is now being attracted by $-q$. Hence, less work has to be done in transporting it to the surface of the charged positive sphere, and hence, the potential of Q has been lowered.

We don't have to make use of electrostatic induction to reduce the potential of a charged body. Any negative charge placed near $+Q$ will lower its potential. A configuration of bodies which allows one (or more) of the constituent bodies to store large amounts of charge while remaining at fairly low potential is called a **condenser** or **capacitor.**

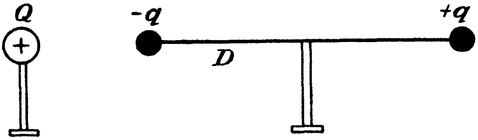


FIG. 15.5.

How shall we measure the effectiveness of a capacitor? We want it to store as high a charge as possible at a given potential. Evidently the ratio Q/V would be an appropriate measure of the electrical storing effectiveness of a condenser. We call this constant ratio the electrostatic **capacitance** or **capacity** and define it as follows:

$$\boxed{C = \frac{Q}{V}} \tag{15.1}$$

$$\text{Unit of capacitance} = \frac{\text{unit of charge}}{\text{unit of potential}}$$

The practical unit is

$$1\ \textbf{farad} = \frac{1\ \text{coulomb}}{1\ \text{volt}} = 10^6\ \text{microfarads}$$

$$1\ \text{farad} = 9 \times 10^{11}\ \text{esu of capacitance}$$

$$1\ \text{esu of capacitance} = \frac{1\ \text{statcoulomb}}{1\ \text{statvolt}} = 1\ \text{cm}$$

It is left as an exercise to the student to verify these relationships between the units.

Example 1. Capacitance of a Sphere

Let us determine the capacitance of a sphere of radius $R = 10$ cm in esu.

Solution. The constancy of the ratio $C = Q/V$ follows from Coulomb's law. The work done in moving a unit charge from infinity to the surface of the sphere is proportional to the coulomb repulsion at any point of the path and hence to the charge Q toward which the unit charge is moved. This is expressed by writing $V = kQ$, where k is a proportionality constant. This is identical with the equation $V = (1/C)Q$ (or $C = Q/V$), where the proportionality constant is $1/C = k$.

Our objective is now to find an expression for V in terms of Q, expressing the work done against the coulomb force in moving a unit test charge toward the sphere. Substituting this expression in $C = Q/V$, we expect Q to cancel out from the numerator and denominator, yielding the desired expression for C. The required expression for V follows from the following preliminary consideration:

Potential Difference in a Coulomb Field

Imagine two points P_0 and P_n at the distances r_0 and r_n from a positive point charge Q (Fig. 15.6). We undertake to determine the potential difference V between the points P_0 and P_n, that is, the work done by the Coulomb force on $q = +1$ esu of charge as it moves from P_0 to P_n.

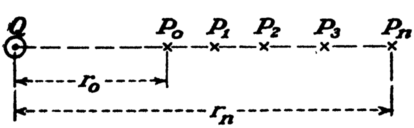


FIG. 15.6.

We imagine the path P_0P_n to be subdivided by points $P_1, P_2, P_3, \ldots$ (which are the end points of the radii r_1, r_2, r_3, . . .) into short paths $\Delta r_1 = r_1 - r_0$; $\Delta r_2 = r_2 - r_1$; . . . ; $\Delta r_n = r_n - r_{n-1}$. If we assume Δr to be small enough, the Coulomb force Q/r^2 exerted upon a unit charge can be assumed practically not to vary over the interval Δr, and we can express the amount of work done in moving the unit charge from P_0 to P_1 by the following approximate expression:

$$\Delta V_1 = F_1\,\Delta s_1 = \frac{Q}{r_1^2}\,\Delta r_1 \tag{15.2}$$

Similarly we obtain for other path elements

$$\Delta V_2 = \frac{Q}{r_2^2}\,\Delta r_2; \qquad \Delta V_3 = \frac{Q}{r_3^2}\,\Delta r_3; \qquad \ldots; \qquad \Delta V_n = \frac{Q}{r_n^2}\,\Delta r_n \tag{15.3}$$

The total amount of work done over the path P_0P_n is

$$V = \sum \Delta V = \frac{Q}{r_1^2}\,\Delta r_1 + \frac{Q}{r_2^2}\,\Delta r_2 + \cdots + \frac{Q}{r_n^2}\,\Delta r_n = Q \sum \frac{\Delta r}{r^2} \tag{15.4}$$

To evaluate this sum we use a simple mathematical artifice. We introduce a new variable

$$z = \frac{1}{r} \tag{a}$$

and define

$$\Delta z_1 = z_0 - z_1 = \frac{1}{r_0} - \frac{1}{r_1} = \frac{r_1 - r_0}{r_1 r_0} = \frac{\Delta r_1}{r_1 r_0}$$

If we choose Δr_1 small enough, we can make the difference between r_1 and r_0 as small as we please. We can write then approximately $r_1 r_0 \cong r_1^2$ and obtain for Δz_1 [making use of equation (*a*)]

$$\Delta z_1 = \frac{\Delta r_1}{r_1^2} = \Delta r_1 z_1^2 \tag{b}$$

or

$$\Delta r_1 = \frac{\Delta z_1}{z_1^2} \tag{c}$$

Similarly we obtain

$$\Delta r_2 = \frac{\Delta z_2}{z_2^2}; \qquad \Delta r_3 = \frac{\Delta z_3}{z_3^2}; \qquad \ldots ; \qquad \Delta r_n = \frac{\Delta z_n}{z_n^2}$$

Substituting in equation (15.4) these equations for Δr and $r = 1/z$ from equation (*a*), we obtain

$$\begin{aligned} V &= Q \sum \frac{\Delta r}{r^2} = Q \sum \frac{\Delta z/z^2}{1/z^2} = Q \sum \Delta z = Q(\Delta z_1 + \Delta z_2 + \Delta z_3 + \cdots + \Delta z_n) \\ &= Q[(z_0 - z_1) + (z_1 - z_2) + (z_2 - z_3) + \cdots + (z_{n-1} - z_n)] \end{aligned}$$

We see that only the terms z_0 and z_n occur once. All the others such as z_1, z_2, z_3, etc., occur twice, once with the positive and once with the negative sign in adjacent parentheses and, hence, cancel out. We obtain thus

$$V = Q(z_0 - z_n) \tag{15.5}$$

Now we substitute the values of z from the definition equation (*a*) and get

$$\boxed{V = Q\left(\frac{1}{r_0} - \frac{1}{r_n}\right)} \tag{15.6}$$

This important equation expresses the potential difference V between two points in a Coulomb field.

Now suppose that P_n is an infinitely distant point. Then $r_n = \infty$, and we get for the **potential at the point P_0 relative to infinity**

$$\underline{V} = Q\left(\frac{1}{r_0} - \frac{1}{\infty}\right) = \underline{\frac{Q}{r_0}} \tag{15.7}$$

Since the charge on a sphere acts upon outside charges as if it were concentrated at the center of the sphere, the potential of the charge of a sphere of radius R is $V = Q/R$. It is the work done in moving a unit positive charge from infinity to the surface of the sphere.

Substituting the value of V from equation (15.7) in equation (15.1), we obtain for the capacitance of a sphere of radius R

$$\boxed{C = \frac{Q}{Q/R} = R} \tag{15.8}$$

The capacitance of a sphere is equal to its radius.
In the electrostatic cgs system the unit of capacitance is 1 *cm. It is the capacitance of a sphere of* 1 *cm radius.* In our problem $\underline{C = 10 \text{ cm.}}$

Example 2

Find the capacitance of a parallel-plate condenser (Fig. 15.7) whose plates have an area of $a = 100$ cm^2 and are separated by an air gap of $d = 0.01$ cm.

Solution. Let us consider for a moment a different type of condenser. Imagine that we have a spherical shell of radius R_1 (Fig. 15.8) carrying a charge $+Q$. The potential at the surface of this sphere will be $V_1 = Q/R_1$. We can reduce this potential by surrounding this sphere with another spherical shell of radius R_2 carrying the charge $-Q$. Now, a unit positive charge moving from infinity will not experience a force of repulsion until it reaches the surface of the larger sphere. This is due to the fact that the charges $+Q$ and $-Q$ act upon points beyond the distance R_2 as if the charges of both spheres were located at the center, where they add up to zero. But work would be done in moving $q = +1$ esu from R_2 to R_1. As soon as the unit charge enters inside the larger spherical space, the charge on the surface of the larger sphere exerts no forces upon it and only the force field of the smaller sphere is effective. The work done is simply equal to the change of potential in the field of the smaller sphere for a displacement from distance R_2 to distance R_1 from the center. This potential difference is given by equation (15.6):

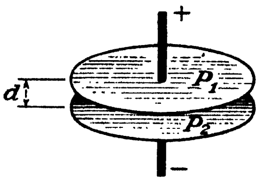


FIG. 15.7. Parallel-plate capacitor.

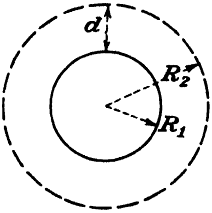


FIG. 15.8.

$$V = Q\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{15.9}$$

This is smaller than the potential of charge Q on sphere of radius R_1 before it is surrounded by the second charged spherical shell.

This device for storing charge at a reduced potential is a so-called *spherical condenser.* The capacitance is by definition

$$C = \frac{Q}{V} = \frac{Q}{Q\left(\frac{1}{R_1} - \frac{1}{R_2}\right)} = \frac{1}{\left(\frac{1}{R_1} - \frac{1}{R_2}\right)} = \frac{R_1R_2}{R_2 - R_1}$$

If the difference $d = R_2 - R_1$ is very small as compared with R_1, $R_2 \simeq R_1 = R$ and we can write approximately

$$C = \frac{R^2}{d} \tag{15.10}$$

or expressed differently,

$$C = \frac{4\pi R^2}{4\pi d} \tag{15.10a}$$

Now $A = 4\pi R^2$ is the area of the sphere of radius R. Hence,

Capacitance of a spherical condenser: $$\boxed{C = \frac{A}{4\pi d}} \tag{15.10b}$$

C is given in centimeters if A is measured in square centimeters and d in centimeters.

It can be shown theoretically and experimentally that equation (15.10*b*) still holds with a good degree of approximation for the capacitance of a condenser in which the two spheres have been replaced by two hemispheres, quarter spheres, or any small fraction of a sphere provided the linear dimension of the area of the segments is large as compared with the distance d. Now a plane plate can be considered as a small area cut out from a sphere of very large radius. Thus a parallel-plate condenser (Fig. 15.7) can be considered as a small segment of a spherical condenser of large radius of curvature. **Thus, equation (15.10b) is applicable to the parallel-plate capacitor.**

In our problem we obtain

$$C = \frac{A}{4\pi d} = \frac{100}{4\pi \times 0.01} = \underline{796 \text{ cm}}$$

The Earth as a Capacitor

The earth, having a huge radius, has a very large capacitance, so that charges which are led off to ground in usual laboratory experiments practically do not alter the potential of the earth. For all practical purposes we are justified in regarding the potential of the earth as constant and might as well use it as our reference level instead of a point at infinity. Thus, *we adopt the convention of calling the potential of the earth zero.* Instead of defining as potential the amount of work required to move a unit positive charge from infinity to a given point, we replace the infinitely distant point for practical purposes by a point on the surface of the earth. In other words, we measure the potential difference against ground. Practically, we are interested only in potential *differences*, so that the choice of reference level is of no importance.

Effect of Dielectrics on Capacitance

If we take a condenser out of a vacuum and immerse it in a dielectric medium such as, for instance, oil, we find an increase in its capacity. The original capacity C is increased to a new value $C' = KC$, where K turns out to be identical with the "dielectric constant" which determines in Coulomb's law [equation (14.1*b*)] how much the electrostatic force between two charges is reduced in the dielectric medium. This suggests a simple way of measuring the dielectric constant of a medium by determining how it increases the capacitance of a condenser when it is introduced into the gap between its plates or, better, when the capacitor is submerged into the medium.

Why is the capacitance changed by a dielectric? The clue is provided by the fact that the dielectric constant occurs in the expression of Coulomb's law. The *increase* of capacitance must have something to do with the *decrease* of the Coulomb force in the medium.

Let us consider the simplest case of the isolated sphere whose capacitance is increased when it is submerged into a dielectric ocean. From equation (15.1) ($V = Q/C$) we see that the *increase* in capacitance implies a *decrease* in potential of the sphere. This means that less work must be done in moving a like charge from infinity to the surface of the sphere in the dielectric medium than in a vacuum. This is no surprise to us. It is a necessary consequence of Coulomb's law [equation (14.16)], which states that the repulsive force is reduced by a factor of K at every point of the path of the test charge. Thus the increase in capacitance is the consequence of the decrease in potential which is due to the decrease of the Coulomb force.

Explanation of the Effect of Dielectrics upon the Electric Field

We stated in connection with the discussion of Coulomb's law that the only action through which a charge reveals itself to us is the force which it exerts upon another charge. We defined a charge as being proportional to the measured force of interaction. Now we notice that the force of interaction between two charges drops as soon as we surround them with a dielectric medium. Similarly, when we place a charge q between the oppositely charged plates of a parallel-plate condenser (Fig. 15.7), a force is exerted upon the charge: $F = Eq = (V/d)q$ (where E = the field intensity in the condenser gap, V = the potential difference between the plates, d = the plate distance, and V/d = the potential gradient). When we fill the gap with a dielectric, the force drops to the value of $F = (V/K)(q/d)$. Throughout this entire process there was no leakage of charge from the condenser, and yet the only way we could imagine such a reduction in force to come about is by reduction in the charge on the condenser plates. How can we imagine the charge to diminish without leaking off? Let us make the following hypothesis: We assume the dielectric to contain **electric dipoles.** They may be created when the dielectric is exposed to an electric field by electrostatic induction in tiny particles present in the dielectric within which electric charges can be shifted (Fig. 15.9*a*), or they may be present as so-called *permanent dipoles* in the dielectric and may be merely oriented by the field (Fig. 15.9*b*). We have encompassed by dotted squares adjacent opposite charges of the dipoles which are so close to each other that they neutralize each other in interaction with extraneous charges. The only charges which are not compensated by opposite charges of adjacent dipoles are the negative charges in the uppermost layer of the top dipoles and the positive charges in the lowermost layer of the bottom dipoles. These charges are free to compensate part of the charges $+Q$ and $-Q$ which were placed on the condenser plates. *Thus we see that the assumption of oriented dipoles in the dielectric is capable of accounting for the observed diminution of the electric field between the condenser plates by showing how the action of the original charges can be partly neutralized by charges of dipoles adjacent to the plates.*

The reduction of the Coulomb force can be explained similarly. Consider the interaction between the charges $+Q$ and $+q$. We can regard $+Q$ as the charge which produces in a vacuum the field $E = Q/R^2$ at the point where q is located. In a dielectric medium the dipoles will orient themselves in such a fashion as to form a layer of uncompensated negative charges $(-Q')$ adjacent to the surface of the body carrying the charge $+Q$. As a result the charge q will not be acted upon by a field created by charge $+Q$ but rather by a field created by the charge $Q - Q'$.

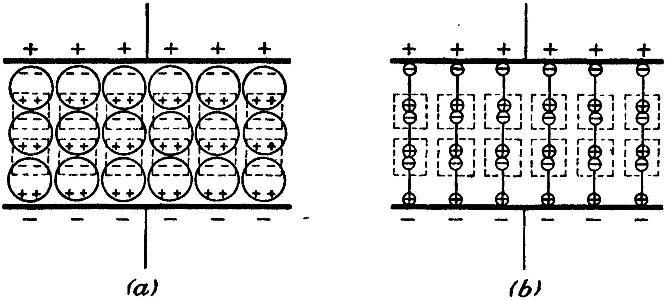


FIG. 15.9. *a.* Induced dipoles in the homogeneous electric field of a parallel-plate capacitor. *b.* Orientation of permanent electric dipoles in the uniform field of a parallel-plate condenser.

The process of orientation or induction of dipoles is referred to as **polarization of the dielectric.**

Comparison between Magnetization and Dielectric Polarization

When we place a piece of metal in a nonhomogeneous electric field (*i.e.*, one whose intensity varies in space), such as the Coulomb field of a point charge Q, the piece of metal is attracted toward Q. We explained it by separation of charges in the metal body by the electric field and by preponderance of the force of attraction exerted upon the induced opposite charge. A similar phenomenon of attraction is observed when dielectric particles are exposed to a nonhomogeneous electric field, and the conditions for attraction of iron objects in the nonuniform field of a magnet are quite analogous.

Can we conclude from such observations that there is a separation of charges in a dielectric body similar to the induction in a conductor and that a separation of "south magnetic and north magnetic fluids" occurs in a piece of iron or nickel exposed to a magnetic field? We shall see that this conclusion is discouraged by further experiments which show that the analogy between the latter two cases and electrostatic induction is quite superficial. Let us imagine the body B in Fig. 15.1 to be made of dielectric material. We find that B is indeed attracted toward R but much less strongly than a comparable piece of metal. The ends of B_1 and B_2 behave as if they carried a negative and a positive charge, respectively, which is similar to results obtained with a metallic body B. But when we separate B_1 and B_2 and remove the charged rod R, we find no unbalanced charge on either of the bodies. Hence, the process is essentially different from electrostatic induction in metals. If we test B_1 and B_2 *after separation* but with body R *still in position*, B_1 and B_2 display each a positive and a negative electrification at both ends. The charges are disposed in them in the same manner as in the body B before separation. This inability on our part to separate the positive and negative charges in the dielectric is very similar to the impossibility of isolating the north pole from the south pole of a magnet by breaking it in two.

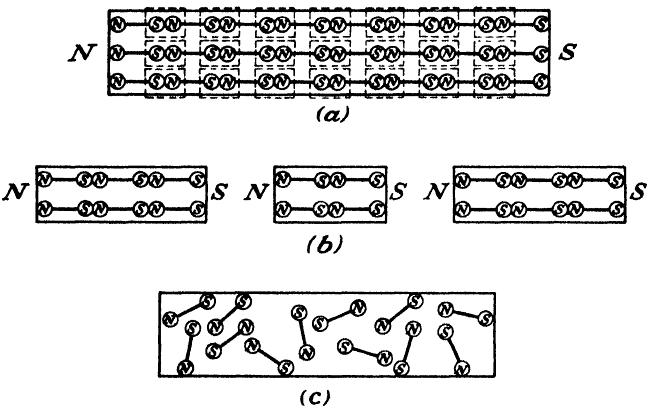


FIG. 15.10. *a*. Orientation of elementary magnetic dipoles in a magnetized iron bar. The poles at the ends are formed by the uncompensated poles of the elementary magnets. The poles in the interior which compensate each other are surrounded by dotted squares. *b*. The same bar is cut in two places. We obtain again complete magnets with two poles each. *c*. Unmagnetized iron bar. The elementary magnetic dipoles are oriented at random.

It seems that the same kind of hypothesis would explain both phenomena. This is indeed the case. According to the polarization hypothesis of the previous paragraph, it is impossible to obtain a surplus of charge in one of two pieces into which a polarized dielectric bar is broken if one assumes that the elementary dipoles are not breakable. Similarly, if we imagine a magnetized iron bar to differ from an unmagnetized bar by the fact that the former contains oriented magnetic dipoles (whereas they are oriented at random in the latter), we see (Fig. 15.10) that breaking a magnet bar into pieces always yields magnets with two poles.

According to the above hypothesis, the process of magnetization consists of orientation of the elementary magnetic dipoles, or the so-called *elementary magnets*, by a magnetic field to which they are exposed. Jarring the bar aids in the process of orientation. If friction prevents rapid loss of orientation of the elementary magnets, we obtain a permanent magnet.

Behavior of Matter in Magnetic Fields

There is no object which could not be electrified by friction, no substance which would not experience a force when placed in a nonhomogeneous electric field. All matter is electrifiable. Why should the manifestation of magnetic phenomena be limited to a small group of substances such as iron, cobalt, nickel, and some special alloys? This question must have stimulated Faraday to place all kinds of matter into a strong magnetic field in a fashion which would allow the observation of very minute forces. He found that magnetizability is as general a property of matter as electrifiability. Before Faraday's experiments, people considered only iron, cobalt, and nickel as magnetizable, because the forces which they experience in a magnetic field are thousands of times larger than forces exerted upon other materials, so that the magnetization of other kinds of matter can be observed only with very sensitive means employed for the first time by Faraday. Faraday suspended small horizontal rods of various materials and small glass tubes filled with various kinds of fluids between the poles of a strong magnet (Fig. 15.11). The figure shows the top view of the arrangement. The suspension fiber is perpendicular to the plane of the paper and is attached at the point P. The suspended bar forms originally an angle of 45° with the lines of force. Iron, cobalt, and nickel rods are oriented like iron filings with their longitudinal axis parallel to the field lines (Fig. 15.11*B*). Many other substances, such as aluminum, cupric chloride, or oxygen† (in a suitable container), are oriented similarly but under the action of much smaller torque. Substances which show this type of orientation are called **paramagnetic.** Substances for which the orienting torque is exceptionally large (iron, cobalt, nickel) are called **ferromagnetic.**

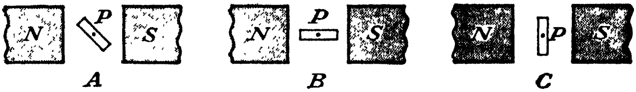


FIG. 15.11. Top view of bars suspended in the gap of a magnet. *A*. Initial position of the bar. *B*. Paramagnetic or ferromagnetic bar. *C*. Diamagnetic bar.

---

† The strong paramagnetism of oxygen has been used as a basis for quantitative chemical detection of small amounts of this substance.

But Faraday found that not all materials were oriented longitudinally. Some rods were oriented at right angles to the lines of force, as shown in Fig. 15.11*C*. He called such substances **diamagnetic.** He thus discovered two new types of magnetism, *paramagnetism* and *diamagnetism.* A paramagnetic body is attracted toward a magnetic pole *N* (Fig. 15.12*A*), which suggests that a south pole is induced near the north pole of the magnet. A diamagnetic body, however, is repelled from the magnet pole *N* (Fig. 15.12*B*), suggesting that a north pole is induced in it near the north pole of the magnet. How can this strange magnetization of diamagnetic substances be explained? The theory of elementary magnetic dipoles fails to explain it. The explanation will be given on an entirely different basis in Chap. 24.

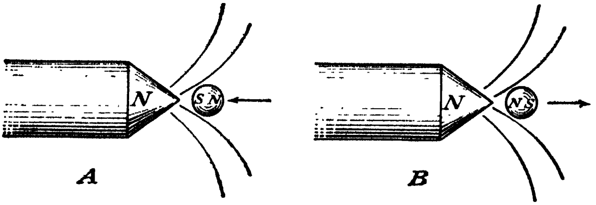


FIG. 15.12. *A*. A paramagnetic sphere is attracted toward pole. *B*. A diamagnetic sphere is repelled from pole.

QUESTIONS AND PROBLEMS

1. *a*. Explain electrical attraction of neutral conductors by a charge.

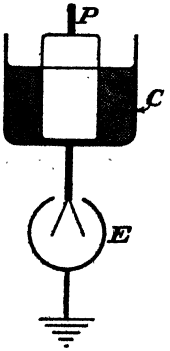
b. In what kind of a field will a neutral body *not* be electrically attracted?

2. If you charge a small metal sphere and touch it to a large uncharged metal sphere, (*a*) does the larger one charge up? (*b*) does the smaller one lose its entire charge? (*c*) If both remain charged, which one will have the larger charge? (*d*) If the larger sphere is hollow and has a hole large enough to pass the smaller sphere through it into the interior of the large sphere, what will happen to the charge of the small sphere when the latter is touched to the inside wall of the large sphere?

3. Occasionally people suffer a severe electric shock without being struck by lightning when a charged thundercloud directly overhead discharges through lightning to a neighboring cloud of opposite charge ("return shock"). Explain this phenomenon. (Hint: Electrostatic induction is involved.)

4. How would you demonstrate, using an electroscope, that the opposite charges induced on the surfaces of B_1 and B_2 in Fig. 15.1 are equal to each other?

5. You are given a charged electroscope. How could you determine in a simple way the sign of the charge?



6. Explain what happens electrically in the following experiment: A can *C* is mounted on an electroscope *E* as shown in the figure. The can is filled with water. To begin with, the electroscope leaves are collapsed, indicating zero charge. Then a paraffin cylinder P is immersed into the water. The leaves do not move. When the block P is withdrawn, the leaves diverge strongly. When the cylinder P is reimmersed, the leaves collapse to zero deflection. Now we withdraw P again and discharge the electroscope so that it reads zero. When we reimmerse P, the electroscope leaves diverge again exactly as much as before. Do you expect the charge of the electroscope to be the same as it was in the preceding experiment? What general principle does this experiment demonstrate?

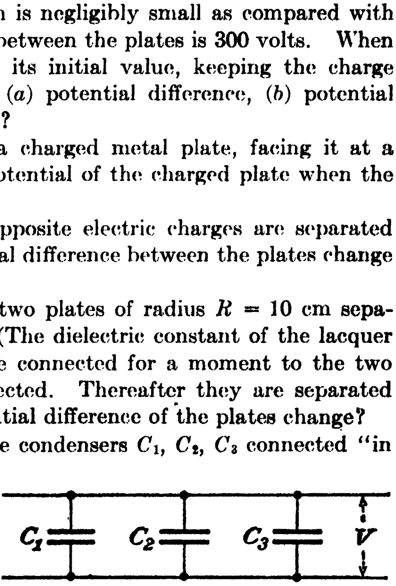
7. The circular plates of a plate condenser, which are charged and insulated from each other, are separated by a distance which is negligibly small as compared with the plate diameter. The potential difference between the plates is 300 volts. When the plate distance is reduced to one-third of its initial value, keeping the charge constant, what is the value of the resulting (*a*) potential difference, (*b*) potential gradient, (*c*) field intensity between the plates?

8. A grounded metal plate is parallel to a charged metal plate, facing it at a distance of 0.1 mm. What happens to the potential of the charged plate when the grounded plate is removed?

9. Two metal plates carrying equal and opposite electric charges are separated by an air gap of 1 mm. How does the potential difference between the plates change when a sheet of mica is placed between them?

10. A parallel-plate condenser consists of two plates of radius $R = 10$ cm separated by a layer of lacquer 0.01 mm thick. (The dielectric constant of the lacquer is $K = 4$.) The two well-insulated plates are connected for a moment to the two poles of a 1.5-volt battery and then disconnected. Thereafter they are separated by a distance of 10 mm. How does the potential difference of the plates change?

★11. Show that the total capacity C of three condensers C_1, C_2, C_3 connected "in parallel" as shown in the drawing is $C = C_1 + C_2 + C_3$.



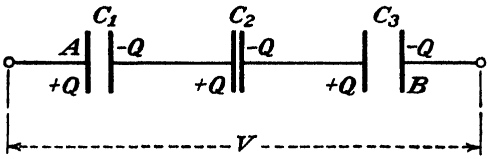
a. How large is the charge Q of each condenser in terms of V and its capacity?

b. Express the total charge on all condensers in terms of V and C.

c. What can you conclude from the answer to *b* about the resultant capacity $C = Q/V$ of the system?

d. Generalize your result for an unlimited number of condensers.

★12. Show that the resultant capacitance C of the system of three condensers C_1, C_2, C_3 connected "in series," as shown in the figure, is given by the expression $1/C = (1/C_1) + (1/C_2) + (1/C_3)$.



a. Explain why the same charge Q appears on all condenser plates with alternate signs (as shown in the diagram) when a charge $+Q$ is applied to plate A.

b. Express the potential difference across each condenser in terms of its charge Q (which is the same for all condensers) and its capacitance.

c. Express the potential difference between the plates A and B in terms of the potential differences of the individual condensers.

d. What conclusion can you draw from the result of *c* about the value of the resultant capacitance $C = Q/V$?

e. Generalize your result to an unlimited number of condensers.

★13. Design a voltage-multiplying circuit consisting of four capacitors which are charged while connected in parallel to a p.d. of 250 volts each. How would you have to change the connections of the condensers in order to obtain 1,000 volts between two terminals of your circuit? Suggest how this could be arranged practically.

14. In making a "tubular condenser," two thin ribbons of aluminum foil 6 cm wide are wound on a spool. Between these ribbons is sandwiched a thin ribbon of dielectric material 0.2 mm in thickness and 8 cm wide, having a dielectric constant $K = 4.0$. How long must the ribbons be if the capacitance of the condenser is to be 0.01 microfarad ($= 10^{-8}$ farad)?

★15. A charge of 10 esu placed on a metal sphere of 1 cm radius is immersed in oil of dielectric constant $K = 2.5$. A solid dielectric sphere of 2 cm radius made of a material of dielectric constant $K = 2.5$ is immersed into the oil so that the centers of the two spheres are separated by $d = 4$ cm. How large a force do the two bodies exert upon each other?

16. Two spheres S_1 and S_2 of radius $R = 1$ cm carrying charges of $Q_1 = 50$ esu and $Q_2 = 10$ esu are placed so that their centers are 10 cm apart in a liquid of dielectric constant $K = 3$.

a. Find the force of interaction between the spheres.

b. Find the field intensity at the center of sphere S_2.

c. Find what the potential at the center of the sphere S_2 would be if only the sphere S_1 were charged.

d. Find the potential at the center of the sphere S_2 considering that both spheres are charged.

e. What is the difference between the values of the potential at the surface and at the center of the sphere S_2?

17. Find the field intensity in dynes per esu of charge between the plates of a parallel-plate condenser 10 cm² in area, 1 mm apart in a vacuum. The p.d. between the plates is 100 volts. How large a force is exerted upon a charge of 4.8×10^{-10} esu between the plates?

18. A charge $Q = +20$ esu is placed at the center of a soap bubble of radius 30 cm. The bubble is assumed to be a conductor.

a. What is the magnitude of the charge separated by induction on the surface of the bubble?

b. Where does the positive induced charge appear?

c. The charge Q is moved radially 15 cm off center. How large is the induced charge on the inner surface of the bubble in this case?

19. The potential at the surface of a charged hollow sphere ($R = 10$ cm) is 50 ergs/esu. The potential in the interior is constant and is the same as at the surface. How much work must be done to move a charge of 20 esu from infinity to the center of the sphere?

20. Describe how you would charge an electroscope (*a*) positively with a positively charged body, (*b*) negatively with a positively charged body. (Give a description and a theoretical account of the process.)

21. Two metal spheres A and B are of equal size and are separated by a distance of twice the diameter of the spheres. If an interaction between the two spheres is known to be due to electric charges, are you justified in concluding that (*a*) both spheres are charged with electricity of the same sign if they repel each other? (*b*) the spheres are necessarily charged oppositely when they attract each other? (*c*) only one of the spheres is charged when they attract each other?

CHAPTER 16

THE CONCEPTS OF TEMPERATURE AND OF QUANTITY OF HEAT

a. In container *A* you have a gallon of boiling water; in container *B* you have a quart of boiling water; in container *C* you have a teaspoonful of boiling oil. (1) Which container required the burning of the largest amount of gas to be heated? which the next largest? (2) Which is hotter, *A* or *B*? (3) Which is hotter, *B* or *C*?

b. Why is our skin sensation unreliable in gauging temperature?

c. How can we decide whether two bodies are equally hot?

d. What is temperature?

e. How can we measure it?

f. Is the definition of temperature given in this text adequate for scientific purposes?

g. Do all bodies expand when heated?

h. Is water suitable as a thermometric fluid?

i. Which are the advantages and disadvantages of the following types of materials as thermometric substances: (1) gases, (2) liquids, (3) solids?

j. Is the reading of a mercury thermometer determined by the expansion of the mercury alone?

k. What is meant by "absolute zero"?

l. How do you compute the absolute temperature from the centigrade temperature?

m. What questionable procedure led to the original conception of the absolute zero?

n. How do we measure quantity of heat? Do we make use of the temperature concept in this process?

o. How do we define specific heat?

p. On what factors does the thermal capacity of a body depend?

q. On what general physical principle is the calorimetric method of mixtures based?

The Intuitive Concept of Heat

In Chap. 15 we encountered the example of a physical quantity which required for its description, in addition to a quantity measure, an intensity measure. We saw that the electrical behavior of a charged body is not determined completely by the magnitude of its charge but that its electrical potential is of equal interest to us.

A similar situation is encountered in the study of heat. In addition to the less familiar concept of the quantity of heat we shall introduce an intensity measure, the **temperature,** of which we have an intuitive conception.

The terms hot and cold are vague expressions describing a sensation evoked by contact of objects with our skin. Whether we find an object hot or cold depends not only on the state of the object but also on what we touched prior to our contact with the object and on the nature of the material. For instance, if we dip our hands into a bath of tap water after withdrawing them from a bath of molten paraffin, we find the tap water cold. If we had our hands previously in ice water, we should find the tap water warm. Or if we have three objects in an oven, two brass cylinders and a wooden one, the two brass cylinders will feel equally hot while the wooden cylinder feels cooler. And yet we shall see presently that all three bodies are equally hot according to physical criteria.

Thus we cannot rely on our sensation even in merely deciding whether or not two bodies are equally hot, and an objective method of measuring "hotness" appears desirable.

Thermometry

The selection of an appropriate method is suggested by observation. We notice that, as we feel a body grow hotter or colder, some of its physical properties, such as volume, hardness, electrical properties, and sometimes color, are seen to change simultaneously. It is then a natural step to choose a body which displays a conspicuous and reproducible change, say, for instance, in volume, as it is heated and to use its volume instead of our skin sensation as a criterion of the "hotness" which we shall associate with a physical property of the body, the **temperature.** Whenever we see the body expand, we shall know it grows hotter (*i.e.*, its temperature rises), and whenever we see it resume its original size, we shall know it has come back to its original temperature.

But such a temperature meter, or, as we call it, a **thermometer,** is not very useful if it indicates merely its own temperature. Suppose we want to measure the temperature of a body B; can we use our thermometer C for it? This question is answered by the observation that two originally unequally hot bodies acquire eventually the same temperature when brought into intimate contact. *We shall define the temperature of a body B as the temperature of the body C if the latter has been in contact with the former for a very long time.*

If the initial temperature of the thermometer is different from the temperature of the body, the final temperature will not be the same as the original temperature of the body. In order not to change the temperature of the body excessively by the process of measurement, the thermometer should be as small as possible. (This is an illustration of how we usually disturb by the process of measurement the quantity which we measure.)

It is not sufficient for a thermometer merely to help us decide whether or not two bodies are equally hot. We should like it to indicate the "degree of hotness" of a body in some reproducible and unambiguous fashion. This can be done, at least theoretically, as follows: There are materials like invar steel which practically don't change in length as they are cooled or heated. (How would you prove that?) We prepare a ruler out of such a material and use it in connection with, say, a brass rod, whose length can now be measured as it is brought in contact with various hot and cold objects. It is a known fact that the melting and boiling (at a fixed pressure) of substances usually takes place at the same constant, reproducible "temperature," which is peculiar to each substance. We can select a variety of such substances and *use their melting points and boiling points* **as fixed points** for our brass-rod thermometer. We could then mark the particular lengths of our brass rod, such as melting point of ice, boiling point of ether, boiling point of alcohol, boiling point of water, or melting point of sulfur. This would give us a thermometer by means of which we could assure the reproduction of a certain "degree of hotness" or gauge the "hotness" of a given body.

But this system of assigning names to various "degrees of hotness" is too clumsy. It is analogous to the designation of railroad stations along a train line. How shall we name and identify intermediate points? This problem is easily solved by choosing numbers as our names for the various "stations." We can then subdivide and designate by appropriate numbers a dense sequence of intermediate thermometer readings. **The number so assigned to our thermometer reading we define as the temperature** of our thermometer. This definition of temperature seems at first glance unsatisfactory because of its arbitrariness. It leaves us completely in the dark about the "true meaning" of temperature. The only value of this definition consists in its enabling us (using the same thermometer) to reproduce or recognize a certain "degree of hotness."

Liquid as a Thermometric Substance. Having adopted a principle of thermometry, let us choose now a better thermometric substance than a brass rod. Liquids usually expand more than solids, and in addition, the expanding liquid can be made to enter a narrow "capillary" glass tube, thus giving large linear excursions for small volume changes.

This kind of a thermometer (Fig. 16.1) measures the difference between the expansion of glass and of the liquid. As it is immersed into a hot bath, the glass expands. This tends to make the liquid level go down, but the expansion of the liquid tends to push the level up, and since the expansion of mercury, for instance, is about seven times as large as that of glass, the net effect is the rise of the mercury column. We can use this expanding mercury column as a thermometer similarly to the way we used the brass rod, but it seems simpler, instead of taking many **fixed points** as previously suggested, to choose only two, say the **melting point of ice and the boiling point of water** (at normal atmospheric pressure), and to divide the elongation of the mercury column corresponding to this temperature change in, say, 100 equal steps, which we shall call **degrees centigrade.**

Choice of a Thermometric Fluid. Of course, we could do the same using quartz instead of glass and alcohol, water, pentane, or some other liquid instead of mercury. If we do this, however, we shall be severely disappointed and confused. If we immerse, say, a mercury-glass and an ether-glass thermometer in the same tub of water (which we stir and assume to be at uniform temperature), then if the mercury thermometer reads 0°, the ether thermometer, too, reads 0°. If the mercury thermometer reads 100°, so will the ether thermometer. (Why?) But if the mercury thermometer reads 50°, the ether thermometer will not read 50°. The mercury and ether thermometers will disagree about all temperatures except 0 and 100°. Which one reads the correct temperature?

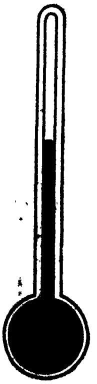


FIG. 16.1. A glass-liquid thermometer.

We shall be even more confused if we immerse these two thermometers with a water-glass thermometer into a bath of gradually rising temperature. Slightly above the freezing point of water the water thermometer will be falling while the other two will be indicating a rising temperature. Does the temperature really rise or fall?

Our confusion is due to the fact that these liquids do not expand alike. If we plot the readings of an ether thermometer as abscissae and those of the mercury thermometer as ordinates, we do not get a straight line. This means that we must be even more specific and arbitrary in defining temperature. We have to specify the thermometric fluid and the material of the solid container.

As long as we are at liberty to choose a thermometric fluid, we might as well choose the most advantageous one. It turns out that gases expand even more than liquids with rising temperature. Whereas the expansion of mercury is 7 times as great as that of glass, the expansion of air exceeds that of glass 146 times. Thus for most practical purposes the expansion of the vessel can be neglected in a gas thermometer. If we compare various gas thermometers (air, hydrogen, helium, oxygen, etc.) by plotting their readings, as before, along two mutually perpendicular axes, we get a straight 45° line, indicating that all gases expand practically alike. Thus the advantage of defining temperature on the basis of a gas thermometer is obvious.

Figure 16.2 shows a standard gas thermometer. A glass bulb *B* is connected through a glass tubing *L*, and the rubber tubing *T* to a glass tube *R*. The system is filled with mercury as shown in the diagram.

Instead of watching the expansion of the gas as it is revealed by migration of the mercury level, it is often more convenient for precise measurements to use such a device as a "constant-volume thermometer." In this case the leg R is raised until the mercury level in leg L reaches the mark M. As the bulb B is heated, the mercury level in L goes down but can be brought back by raising leg R. The mercury level in R can be read on the scale S. The measured difference h between the two mercury levels indicates the excess of the gas pressure in the bulb B over atmospheric pressure. This height is taken as an index of temperature. Atmospheric-pressure variations influence the readings. The value of the atmospheric pressure must be taken into account (see Chap. 18). Measurements are made by immersing B into the container in which the temperature is to be determined.

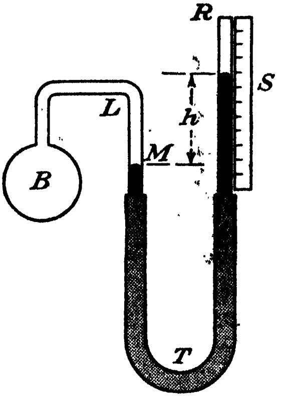


FIG. 16.2. A gas thermometer.

Operational Definition of Temperature

To many beginners who want to know what temperature "really" is, this operational definition of temperature by mere convention, defining it merely as the numerical reading of a gas thermometer, must be a shocking disappointment. However, we shall be in a position to offer an additional *theoretical definition* of temperature when we discuss the kinetic theory of gases. This will satisfy our imagination more than the present operational definition, which defines temperature by merely telling us how we can measure it.

We shall henceforth understand as centigrade temperature the numerical indication of a helium thermometer which is graduated in 100 equal steps (degrees centigrade) between the readings for melting ice (0°C) and boiling water (100°C) (at normal atmospheric pressure). For higher and lower temperatures these steps are extended beyond the two "fixed points."

The Fahrenheit scale, which is used in the United States in medical and engineering thermometry, is defined as follows: The temperature of melting ice is labeled as 32°F, and the temperature of boiling water as 212°F. Thus the temperature interval of 100°C corresponds to an interval of 180 Fahrenheit degrees. From this follows the following two conversion formulas in which F stands for the number of Fahrenheit degrees and C for the corresponding number of centigrade degrees:

$$C^\circ = \tfrac{5}{9}(F^\circ - 32)$$
$$F^\circ = \tfrac{9}{5}C^\circ + 32$$

Thermal Expansion

Linear Expansion. The problems involved in thermometry lead us to the investigation of thermal expansion of solids, liquids, and gases. We find that the elongation ΔL of a solid rod is proportional to the length of the rod at a fixed standard temperature and to the temperature change:

$$\boxed{\Delta L = \alpha L_0 \, \Delta t} \tag{16.1}$$

In this equation $\Delta L = L_2 - L_1$ is the elongation due to a temperature increase $\Delta t = t_2 - t_1$, and L_0 is the length of the rod at 0°. The proportionality constant α is called the **coefficient of linear expansion.** Its physical meaning is seen when we set $L_0 = 1$ unit length and $\Delta t = 1°\text{C}$. Then $\alpha = \Delta L$, that is, the elongation of a unit length per degree centigrade. (The units of length and temperature can be chosen arbitrarily.) Obviously, this equation can also be written so as to express the length of the rod at any temperature:

$$L_2 = L_1 + L_0 \alpha \, \Delta t \tag{16.1a}$$

For small values of Δt, α being usually very small as compared with 1, a satisfactory approximation is obtained for solids by replacing L_0 by L_1, since L_1 is usually quite close to L_0 in magnitude and since a *small change* in the *small quantity* $\alpha L_0 \, \Delta t$ does not introduce a large error in the above equation. This simplifies equation (16.1) as follows:

$$\boxed{L_2 = L_1(1 + \alpha \, \Delta t)} \tag{16.1b}$$

This equation not only applies to the expansion of a rod but holds for the change of distance between any two points of the body regardless of its shape.

Cubical Expansion. A change of all linear dimensions of a solid is accompanied by a volume change, and a similar consideration as above leads to an analogous expression for the volume of a solid:

$$\boxed{V_2 = V_1(1 + \beta \, \Delta t)} \tag{16.2}$$

where β is the so-called **coefficient of volume expansion,** or coefficient of cubical expansion. *There is a simple approximate relationship between* α *and* β: $\beta \cong 3\alpha$. Its proof is left to the student.

The volume expansion of a solid container plays a role in thermometers. A hollow container expands as if it were entirely filled with the material of which it is made. The knowledge of the coefficient of linear expansion of glass enables us to determine the volume expansion of the thermometer bulb, which partly determines the thermometer reading.

One might think of applying equation (16.2) also to the volume expansion of liquids, since the same arguments are valid for the cubical expansion of liquids as of solids. But we must not forget that equation (16.2) is an approximate equation, just as equation (16.1*b*) was obtained by the approximation of setting $V_1 \approx V_0$. This approximation was quite good in the case of solids but is not so satisfactory with liquids, which have a much higher coefficient of expansion. We shall use with liquids the unapproximated equation analogous to equation (16.1*a*):

$$\boxed{V_2 = V_1 + \beta V_0 \, \Delta t} \tag{16.2a}$$

or

$$\boxed{\Delta V = \beta V_0 \, \Delta t} \tag{16.2b}$$

If the temperature corresponding to V_1 happens to be the ice point (0°C), $V_1 = V_0$ and we can write

$$\boxed{V = V_0(1 + \beta t)} \tag{16.2c}$$

where V is the volume at the temperature of t°C and V_0 the ice-point volume. t stands for $\Delta t = t - 0$.

Temperature Coefficients of Volume and Pressure of a Gas

Equation (16.2*c*) can also be applied to gases at constant pressure. It turns out that, as a rule, β is practically the same for all (so-called *ideal*) gases and is roughly equal to $\beta = \frac{1}{273}$.

Interestingly enough, the pressure p of a constant volume of gas increases with temperature according to exactly the same law, so that we have

For a gas at constant pressure: $$\boxed{V = V_0\left(1 + \frac{t}{273}\right)} \tag{16.3a}$$

For a gas at constant volume: $$\boxed{p = p_0\left(1 + \frac{t}{273}\right)} \tag{16.3b}$$

The fact that the same form of law holds for the volume changes as for the pressure changes of a gas, the equality of the value of β in both equations, and the fact that β does not depend on the nature of the gas can hardly be a sequence of coincidences. This gives us fruitful food for thought, which will finally culminate in a satisfactory explanation based on an imaginative theory.

The Absolute Zero

Let us now plot the equation (16.3*b*): $p/p_0 = \frac{1}{273}t + 1$. Figure 16.3 represents the ratio p/p_0 (at constant volume) as a function of the centigrade temperature. From our equation we can see that at $t = 0°\text{C}$, $p = p_0$ (which is true by definition of p_0). The intercept with the t axis is obtained by setting $p/p_0 = 0$ (in other words $p = 0$). Then $t/273 = -1$ and $t = -273°\text{C}$. In other words, the pressure p of a confined volume of gas should reduce to zero at $-273°\text{C}$. This may sound like a perfectly respectable scientific statement at first glance.

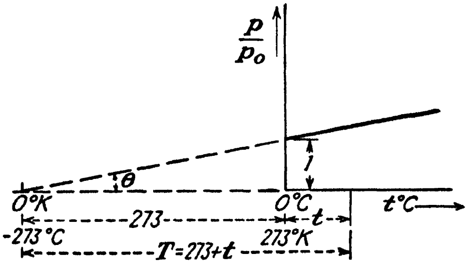


FIG. 16.3. Variation of gas pressure with temperature for an "ideal" gas.

But let us plot the analogous equation (16.3*a*) in the same fashion; *i.e.* we plot V/V_0 as a function of t. The resulting graph is the same as Fig. 16.3 except that p/p_0 is replaced by V/V_0. We get a similar result: The volume of a certain amount of gas maintained at a constant pressure should drop to zero at $-273°\text{C}$. This appears a little more striking and less sensible than the prediction regarding pressure. It is difficult to reconcile this with the law of conservation of mass. If the original mass of the gas cannot vanish, what is the meaning of a mass (of gas) without a volume?

For obvious reasons this temperature of $-273°\text{C}$ (or, more precisely, $-273.2°\text{C}$) has been termed the **absolute zero.**

Fortunately, our general principles are not so seriously imperiled by the above consideration as it may seem at first. As a rule, gases cease to obey the equations (16.3*a*) and (16.3*b*) as they are cooled long before they reach $-273°\text{C}$. Eventually they are liquefied and cease to be a gas. Thus, the above-mentioned critical situation never arises. This consideration demonstrates the *danger of extrapolating a law far beyond the region for which its validity has been established.*

Similarly, we could, for instance, extrapolate the expansion formula for mercury to zero volume [using equation (16.2*c*)] and obtain an "absolute zero" of $-5000°C$ and with other liquids still other values for absolute zero. Does it mean then that the concept of absolute zero has no physical meaning? If it has a sensible meaning, it certainly doesn't follow from our considerations so far. But as we shall see later on, *the temperature of $-273.2°C$ does have a physical significance. It is the lowest temperature which we can conceivably reach.* (It has been approached to within 0.005°C.)

The Absolute-temperature Scale

Kelvin suggested shifting the zero of our temperature scale from the "ice point" to −273.2°. This new scale is called the **absolute scale**† or the **Kelvin scale.**

As can be seen from Fig. 16.3,

$$\boxed{T^{\circ} = t^{\circ} + 273.2}$$

T is the temperature in degrees Kelvin (°K) and t the centigrade reading. The absolute-temperature scale is widely used in physics. T is referred to as the "absolute temperature" of a body.

Quantity of Heat

Having elaborated on the concept of temperature, let us now turn to the question, "What is heat?" We already answered this question in Chap. 10 by recognizing that heat is a form of energy. In the case of heat we must consider, in addition to the measure of the quantity of energy, a measure of intensity level which governs the transfer of energy from one body to another. Heat is transferred spontaneously only from bodies at a higher temperature to those at a lower temperature.

The intensity level of heat, or temperature, depends on the concentration of energy in an homogeneous body as can be seen from the following conceivable experiment. Let us imagine a mechanism which enables us to convert work entirely into heat through some frictional mechanism. Suppose we immerse this heat generator into a quart of water and do 100 ft-lb of work. We shall generate a certain amount of heat, and the temperature of the water will rise. If we repeat this experiment with $\frac{1}{2}$ qt of water, the temperature rise will be about twice as high for the same amount of work done. (It would be almost exactly twice as high if our heat generator were of negligible dimensions as compared with the water mass.) Evidently, the higher temperature in the second case is due to the concentration of the developed heat energy in a smaller mass of substance. By doing more work we can increase the temperature of our water, since this again increases the concentration of heat energy in the substance.

---

† This name is due to the fact that this scale can be defined independently of the thermometric substance as has been shown by Kelvin.

We shall define as a unit of the quantity of heat (1 cal)† that amount of heat energy which is required to raise the temperature of a unit mass (1 gm) of water by 1°C. Since, as it turns out, the energy required to raise the water temperature from 0 to 1°C is not quite the same as in the case of a rise from 98 to 99°C, **we specify in the above definition of a calorie a rise from 14.5 to 15.5°C.** (This value was chosen arbitrarily in the vicinity of a room temperature of about 15°C.)

Other definitions of a heat unit are possible and are in use, for instance, Btu (the British thermal unit) which is used by engineers in English-speaking countries. It is defined as the amount of heat required to raise 1 lb of water from 63 to 64°F. 1 Btu = 252 cal.

It will take, obviously, m calories to raise the temperature of m grams of water by 1°C and $Q = m\,\Delta t$ calories to raise it by Δt°C.

Specific Heat

Let us now immerse our mechanical heat generator into the same mass m of a different liquid (or solid powder or gas). We shall find that the same amount of work as we used in the case of water will now increase the temperature by a different amount. If the temperature rise of an equal mass of water was Δt°, it will now require s times as much work or heat energy to produce the same temperature rise. Our equation for the energy (in calories) required to raise the temperature of m grams of the new substance by Δt°C will be

$$\boxed{Q = sm\Delta t} \tag{16.4}$$

s is called the **specific heat** of the substance. Its physical meaning is easily seen when we write $s = Q/(m\,\Delta t)$. Set $m = 1$ gm and $\Delta t = 1$°C. Then $s = Q$, that is, the **specific heat of a substance is equal to the amount of heat (in calories) that is required to raise the temperature of a 1-gm specimen by 1°C.** The specific heat is not a constant, but its variation over the range between 0 and 100°C is small enough to be neglected in practical problems.

The specific heat of water is thus 1 cal/deg C. It is unusually high as compared with other substances.

---

† One must distinguish between gram calories (calories) and kilogram calories (Calories) according to whether the unit mass chosen in this definition is 1 gm or 1 kg.

The product (sm) in equation (16.4) is called the **thermal capacity** of a body.

$$(sm) = \frac{Q}{\Delta t}$$

In words: *The thermal capacity of a body is the amount of heat required to raise its temperature by* $1°C$. Equation (16.4) states that the change Q in the heat content of a body is equal to its thermal capacity times the temperature change.

Conservation Equation for Heat Energy

We can now apply the law of conservation of energy to a closed system of bodies which are initially at different temperatures.

Suppose we take n masses: $m_1, m_2, m_3, \ldots$, of the specific heats $s_1, s_2, s_3, \ldots$ at the temperatures of $t_1, t_2, t_3, \ldots$. What will be the final equilibrium temperature of the system? The answer is simple: Since the system is closed, it can neither lose heat to its environment nor gain heat from it. Whatever heat is lost by one body will be gained by some of its neighbors. Thus, the sum of all positive and negative changes in the heat content of the bodies (*i.e.*, heat gains and losses) will be zero. Heat exchanges will proceed until all the bodies are at the same temperature. Thus, there will be some final "equilibrium temperature" assumed by the system.

The change in heat content of a body, as we saw, is the thermal capacity times temperature change:

$$\Delta Q_1 = (s_1 m_1)(t_1 - t) = (s_1 m_1)\Delta t$$

where t_1 is the original and t the final equilibrium temperature of the system. Thus we can write the conservation equation for our system:

$$\Delta Q_1 + \Delta Q_2 + \Delta Q_3 + \cdots$$
$$= \boxed{0 = (s_1 m_1)(t_1 - t) + (s_2 m_2)(t_2 - t) + \cdots} \qquad (16.5)$$

Some bodies cool off in this process, namely, those whose original temperature, say t_1, was higher than the final equilibrium temperature t. For them $\Delta t = t_1 - t$ is positive. Others heat up in this process. For these bodies the initial temperature t_2 is below the final temperature t; in these cases $\Delta t = t_2 - t$ is negative.

This conservation equation is the basis of the method of mixtures for measurement of specific heats. A mass m_1 of the substance whose specific heat s_1 is to be measured is heated to a known temperature t_1

(usually temperature of boiling water) and placed in a **calorimeter.** The latter is a metal container of known mass m_2 and specific heat s_2 filled with m_3 grams of water ($s_3 = 1$ cal/gm). The initial temperature t_2 of the calorimeter and water is measured. The instrument is well insulated by a suitable enclosure against heat exchange with the surrounding air. After the test sample m_1 is introduced, the temperature of the system reaches an equilibrium value t after prolonged stirring and waiting. Since all data are known from available tables and from measurements except the value of s_1, the specific heat of the specimen can be determined from equation (16.5).

Conversion of Chemical Energy into Heat

When two substances A and B unite to form a new substance C, energy is either absorbed or evolved in the process. The former type of process is designated as an **endothermic reaction,** whereas processes of the latter kind are called **exothermic reactions.** Combustion is an important example of an exothermic type of reaction. The heat that is liberated in the process of combustion is converted into mechanical energy in steam engines, internal combustion engines (such as the gasoline motor), and some types of explosives. A slow process of combustion (oxidation) also takes place in the body of animals. It is the source of energy required for the maintenance of the functions of the living organism, the most conspicuous of which is the work done by contracting muscles.

The amount of energy which is liberated per gram of a given substance in the process of combustion can be found for various fuels and food stuffs. The following list offers a few examples:

Substance	*Heat of combustion cal/gm*
Coke	7,000
Gasoline	11,500
Hydrogen	34,000
Starch	4,200
Vegetable protein	4,000

Example

In a process of hardening, six steel files weighing 300 gm each are heated to 750°C and then dipped into 10 kg of oil which is kept in a brass container weighing 2 kg. The initial temperature of oil and container is 10°C.

What is the final temperature after the immersion of the files if heat losses to the environment are neglected?

Data. Specific heat of oil: $s_1 = 0.45$ cal/gm.
Specific heat of steel: $s_2 = 0.15$ cal/gm.
Specific heat of brass: $s_3 = 0.09$ cal/gm.

Solution.

Mass, gm. m_n	Specific heat s_n	Thermal capacity $s_n m_n$	Initial temperature, °C t_n
Oil $m_1 = 10^4$	$s_1 = 0.45$	$s_1m_1 = 4{,}500$	10
Steel $m_2 = 6 \times 300 = 1{,}800$	$s_2 = 0.15$	$s_2m_2 = 270$	750
Brass $m_3 = 2{,}000$	$s_3 = 0.09$	$s_3m_3 = 180$	10

According to equation (16.5),

$$(s_1m_1)(t_1 - t) + (s_2m_2)(t_2 - t) + (s_3m_3)(t_3 - t) = 0$$

Substituting our numerical data, we get

$$4{,}500(10 - t) + 270(750 - t) + 180(10 - t) = 0$$

or

$$(10 - t)(4{,}680) + (750 - t)(270) = 0$$

$$46{,}800 - 4{,}680t + 202{,}500 - 270t = 249{,}300 - 4{,}950t = 0$$

Hence,

$$\underline{t} = \frac{249{,}300}{4{,}950} = \underline{50.4°\text{C}}$$

APPENDIX

Some Biological Considerations and Data

1. The human body is maintained at a temperature above its environment. Heller (1757) taught that "animal heat" was due to friction of blood against the walls of the blood vessels. Crawford and Lavoisier and Laplace showed that the animal heat was due to combustion of food.

The constancy of the blood temperature of warm-blooded animals is remarkable. Temperature variations of the surroundings from −53°C in the arctic to +53°C in the tropics scarcely affect it. Birds show the highest body temperature of all animals (42°C). Most mammals have a temperature of about 39°C, man having 37°C. [The "hottest" mammals are the rabbit and fox (40°C); nearest to man are the horse and elephant (37.6°C).]

2. In judging the advantages of warm-blooded animals over cold-blooded ones it is important to consider that the speed of most chemical reactions is about doubled for a temperature increase of 10°C.

3. Cold-blooded animals maintain their body temperature about 1°C above their environment, whereas the temperature of warm-blooded animals exceeds that of their environment by about 20°C. As a result of the small temperature difference between body and environment, the heat losses of the cold-blooded animals are comparatively very small.

4. Hibernating animals behave during hibernation like cold-blooded animals. Their reduced body temperature minimizes their heat losses, enabling them to get along without food for a long time.

5. Life processes may persist at a reduced rate at very low temperatures. Certain low forms of life, like bacteria, may be cooled down to temperatures of liquid air, kept in suspended animation, and revived by restoring them to their normal optimum

temperature. Omelyansky succeeded in 1902 in reviving microbes taken from the nasal cavity of a mammoth (an extinct mammal), which had been discovered in frozen soil in northern Siberia where it had been buried under ice and snow from 15,000 to 20,000 years.

6. The slowing of metabolism at low temperatures has been used clinically in cases such as the following: A patient whose leg had to be amputated because of gangrene had a heart attack before the scheduled operation. Operation would have been too risky before recovery from the attack. A week's recovery period was required. On the other hand, during that week the patient was liable to die as a consequence of the gangrene. The situation was saved by putting the gangrenous leg into an ice packing and thus retarding the process of gangrene until he recovered to a point where the operation could be performed.

7. If our lakes and ponds were to freeze completely in winter, all fish life would be extinct in them. It is an amazing coincidence (?) in nature that water has a peculiar property which prevents such a disaster. Whereas ordinarily a gas, solid, or liquid contracts as you cool it and, hence, acquires a higher density, water does so only until it reaches the temperature of about 4°C, at which it has a maximum density. When cooled below this point, the water begins to expand and its density decreases. Now let us see what happens in a pond in winter: The water masses at the surface lose heat, become denser, and sink to the bottom, making place for new, warmer quantities of water at the surface. This convective process is very efficient in cooling rapidly the content of the pond. It is more efficient than pure conduction would have been. However, as soon as all the pond water has been cooled to 4°C, the water masses at the surface, which are now being cooled below 4°C, do not sink to the bottom any more but remain at the surface, since they are less dense than the rest of the pond water and, hence, tend to float. Thus the convection process is stopped. The surface freezes, and the loss of heat through the surface continues at a slow rate owing to absence of convection, while fish life continues under the crust of ice. For instance, the water temperature in Lake Superior never sinks under 4°C below a depth of 240 ft.

8. Variations of temperature can be continuously observed with a mercury thermometer. In medical applications, however, it is not always convenient to watch the thermometer while it is in place. But removal of the instrument immediately exposes it to the cooler air and, hence, falsifies the reading. To avoid this difficulty the "clinical thermometer" has been designed to serve as a *maximum thermometer.* It reads the highest division reached by the mercury level. This is accomplished by a constriction in the capillary tube next to the bulb. The mercury expanding in the bulb can easily move through the capillary and push the mercury column ahead, but when it contracts, the mercury column is cut in two at the constriction owing to surface tension and the reading is fixed. In a well-stirred liquid bath the clinical thermometer reaches its final reading within a few seconds! But it is usually kept in the patient's mouth for about 3 min or more. This is due to the fact that the thermometer is cool to begin with and cools the tissues. Since there is no convection, only heat conduction and circulation of blood are effective in supplying the heat to the thermometer bulb. To speed up the process one tends to make the mercury bulb as small as possible. Continuous clinical observations of temperature are usually made by means of electrical thermometers, which will be dealt with later.

9. The proof that the animal heat is of chemical origin was attempted early by Crawford (1779) and by Lavoisier and Laplace. Crawford used a calorimeter con-

sisting of a double-walled box. The interspace between the walls was filled with water.

He first burned coal in his calorimeter and found that, for every 100 oz of oxygen consumed in burning carbon, the temperature of the water in his calorimeter would rise 1.93°F. He then placed a guinea pig in this calorimeter box and found that the consumption of the same amount of oxygen raised the water temperature 1.73°F. These measurements were full of errors (can you think of some?). The conclusion reached was that a certain amount of food gives the same amount of heat whether burned in the open air or in an animal organism.

Later, more precise calorimetric measurements by Rubner, carried out with dogs and using a specially designed air calorimeter, proved with high precision the validity of the law of conservation of energy for physiological combustion.

10. The rate of heat loss of an animal depends on the following factors among others:

a. The ratio of surface to volume: The larger the surface through which the heat generated in a given volume can escape the greater will be the heat loss per second. The following table gives comparative data of surface per unit mass for an adult and a child:

	Mass, kg	Area, cm^2	Area per kg, cm^2
Man	60	14,000	230
Child	10	5,000	500

We see that the child is at a considerable disadvantage in regard to heat losses.

The ratio of surface to volume becomes less advantageous as the size of the animal decreases. Freezing animals flock together into a close group, creating an aggregate having a smaller surface area per unit mass than a single animal.

b. There is a temperature drop of about 5°C from the interior of the human body to the skin. The temperature gradient partly determines the rate of loss of heat.

c. The nature of the subcutaneous tissue also determines the heat flow from the body to the environment. Fat is a poor conductor of heat, and obese persons suffer in summer owing to slow heat losses. Their temperature may go up in the summer, and their metabolism may rise to 50 percent above that of a lean person.

Aquatic mammals are protected from rapid losses of heat by adipose tissue.

11. The temperature of the skin is determined by the circulation of blood as well as by the external air temperature. Thus you cannot obtain the correct skin temperature by placing a thermometer against the skin. The thermometer will eliminate the natural condition in which heat is being lost to the environment by convection and radiation, while heat is being supplied from the interior to the skin by conduction and by the circulating blood. The obtained temperature will be too high. Correct skin temperature can be measured by measuring the heat radiation from the skin (see Chap. 22).

The study of skin temperature is of great clinical value, since from it conclusions can be drawn about the blood circulation in the skin or underlying tissues. In certain so-called *peripheral vascular diseases* the skin temperature is reduced owing to impaired circulation. The therapeutic effectiveness of drugs for such ailments can be appraised by observation of the skin temperature.

Alcohol increases the heat losses of our body, although it makes us feel warmer.

The warm feeling is due to the fact that alcohol dilates the blood vessels in our skin and thus the skin temperature does go up. But our blood is now in more intimate contact with the environment, and hence, the heat losses are more rapid.

12. Motion of air can greatly influence human heat losses by blowing away warm air from our skin and constantly replacing it by cool air. Thus a moderate wind of 15 mi/hr at 20°C will cool a naked person more rapidly than exposure to still air at 2°C. (The accompanying effect of evaporation of sweat will be discussed in Chap. 17.)

13. Clothing saves food, since without clothes we should lose more heat and should have to consume more nutritive fuel.

14. Air is a very poor conductor of heat. The effectiveness of woven woolen clothes and blankets is due to the entrapped air. The role of the wool is to prevent convection of air.

15. The following trouble is frequently encountered with peach trees in the United States. Peach buds often start growing prematurely in winter and are subsequently killed by frost. This undesirable phenomenon is due to the fact that the surface of the twigs is a good absorber of solar radiation, so that on clear winter days the temperature of twigs may rise high above that of their environment. Efforts are being made to develop a peach tree whose twigs will not make such good absorbing surfaces.

QUESTIONS AND PROBLEMS

1. Does the operation of an electric fan affect the temperature of a room?

2. Can you suggest methods for raising the human body temperature without the use of drugs?

3. Why do we shiver when we feel cold?

4. Can clothing affect our metabolism?

5. In designing a mercury-glass thermometer using a prescribed amount of mercury, would you prefer to use a spherical bulb to a cylindrical one? Why or why not?

6. One usually places a teaspoon into a glass when pouring hot water into it in order to avoid cracking the glass.

a. Why should a spoon protect the glass from cracking?

b. What features of the shape of the glass favor its cracking?

7. Could a malingerer raise his mouth temperature reading without taking the thermometer out of his mouth?

8. Can you think of some pathological conditions in which you expect the skin temperature to be abnormal?

9. Suggest a method which you consider the best for measuring skin temperature.

10. Can you design a thermostatic control for an incubator, *i.e.*, a device which maintains a constant temperature within the incubator?

11. In selecting a proper material for filling cavities in teeth, to which of the following properties of the prospective material will the dentist pay special attention: (*a*) color, (*b*) specific gravity, (*c*) coefficient of thermal expansion, (*d*) heat conductivity, (*e*) melting point?

12. Can you name some surgical procedures in which high temperatures are utilized?

13. How long would you boil a clinical thermometer in order to sterilize it?

14. What is the distinction between heat and cold?

15. Can cold be transferred from one body to another, just like heat?

16. At what temperature are the readings of a Fahrenheit scale thermometer and a centigrade scale thermometer the same?

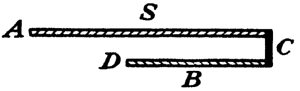
★17. In a certain waterfall water falls from a height of 70 m. Determine the resulting temperature rise of the water.

18. Assume that you have a thermally well-insulated vessel containing 1 kg of water at 0°C and a stirring paddle. If heat losses to the environment can be neglected, how much work would you have to do (convert into foot-pounds) in order to heat the water to the boiling point by stirring? Do you consider such an experiment as practically feasible?

19. What will be the per cent error in measurements with a meter stick made of brass and calibrated at +20°C when it is used at −40°C?

20. A steel ring of 10 cm inner diameter as measured at $t_1 = 5°C$ is heated to $t_2 = 55°C$. Compute (*a*) the change in inner diameter, (*b*) the change in circumference, (*c*) the change in area enclosed by the ring.

21. A steel bar *S* of 50 cm length is rigidly connected to a brass bar *B* of 30 cm length by means of a crosspiece *C*.



a. Compute the change in horizontal distance between the end points *A* and *D* of the bars when the latter are heated from 10 to 60°C.

b. Is it possible to choose the lengths of *S* and *B* so that the horizontal distance between *A* and *D* becomes independent of temperature?

22. How much does the volume of a 2-liter glass flask which was calibrated at 20°C change when it is heated to 200°C?

23. The inner diameter of the glass sphere containing the mercury of a sensitive thermometer is 1 cm. The diameter of the attached capillary is 0.2 mm. Compute the actual excursion of the mercury column per temperature change of 1°C.

24. A gas occupying 2 liters at atmospheric pressure and 0°C is allowed to expand at constant pressure. Find its volume at +273°C.

25. A steel tank contains oxygen at a pressure of 30 atm at a temperature of 0°C. Compute the pressure in atmospheres for $t = 100°C$.

26. A steel tank containing oxygen at a pressure of 20 atm at a temperature of 20°C is heated in a fire to a temperature of 200°C. Compute the rise in pressure in atmospheres.

27. A body of 500 gm mass heated to 100°C is dropped into a calorimeter whose thermal capacity (including the water) is 250 cal/deg C. The initial temperature of the calorimeter is 10°C. The final temperature is 25°C. Find the specific heat of the body.

28. A copper calorimeter ($m_1 = 150$ gm) contains $m_2 = 350$ gm of water at 10°C. A piece of brass of $m_3 = 400$ gm heated to 100°C is placed into the water. Find the equilibrium temperature.

29. A copper calorimeter ($m_1 = 100$ gm) contains 400 gm of water at 20°C. A piece of iron ($m_2 = 300$ gm) heated to 100°C is inserted into the calorimeter. To what temperature do we have to cool a piece of aluminum ($m_3 = 400$ gm) placed into the calorimeter in order that the equilibrium temperature be equal to the initial temperature of the calorimeter?

30. If it were possible to convert the entire heat energy liberated in combustion into mechanical energy, how much gasoline would have to be burned in order to lift the mass of 10 kg to the height of 100 m? (The heat of combustion of gasoline is 11,500 cal/gm.)

CHAPTER 17

TRANSFORMATIONS OF PHASE AND TRANSFER OF HEAT

a. Is it possible to heat a body without raising its temperature?

b. What happens to the energy you spend in rubbing two pieces of wood together? What happens to their temperature?

c. What happens to the temperature of two pieces of ice you rub together? What happens to the energy you spend?

d. What happens to the heat which enters by conduction into a vessel containing ice and water?

e. If you heat boiling water, why is there no rise in temperature?

f. Does water evaporate before the boiling point is reached?

g. What is the difference between boiling and evaporation?

h. What external factor determines the boiling temperature of a given liquid?

i. Does this factor also determine the freezing temperature?

j. What is a saturated vapor?

k. When we try to compress saturated water vapor at constant temperature, what happens to the amount of water which is present in the form of vapor?

l. What happens to the melting point of tin under high pressure? and to the melting point of ice?

m. How can we decide whether pressure will elevate or lower the melting point?

n. How does the human body maintain a temperature below that of its environment in summer?

o. Which property of water prevents the destruction of life by freezing in ponds and rivers in winter?

p. What role do ice and water play in the shaping of the surface of the earth?

q. What is the critical point?

r. What is the critical pressure?

s. Can all gases be liquefied?

t. How does pressure affect the boiling point?

u. How does pressure affect the melting point?

v. How do dissolved substances influence the vapor pressure, boiling point, and freezing point of a liquid?

w. Why does the temperature of an ice-water mixture drop when salt is added to it? Will the same happen if sugar is added instead of salt?

x. On what factors does the rate of heat conduction depend?

y. What is the mechanism of convection of heat?

z. How does the rate of heat radiation depend on temperature?

aa. What does the second law of thermodynamics state?

TRANSFORMATIONS OF PHASE

In Chap. 16 we have shown that the temperature of a body increases as we keep concentrating more and more energy in it by heating it. Is this always the case?

Latent Heat

Let us imagine the following simple experiment: We use again the mechanical heat generator of Chap. 16, *i.e.*, a contraption which converts mechanical energy fully into heat by friction. We place this heat generator in chopped ice and operate it at a constant rate, thus producing heat at a uniform rate.

To begin with we get the expected result (Fig. 17.1). The temperature of the ice increases at a uniform rate from −10 to 0°. But then at 0°C the temperature stops rising for a while in spite of the fact that we continue "pumping" heat into our substance. What is happening to the energy we are investing? Is it being stored in the form of heat despite the fact that there is no temperature rise? So far, the only symptom of a change in the heat content of a body we recognized was a change in its temperature.

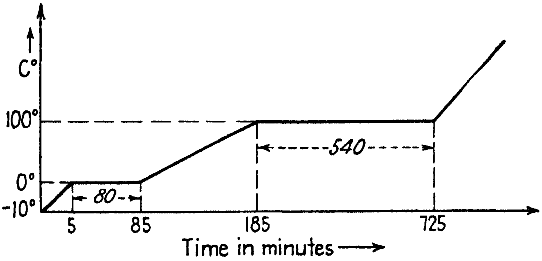


FIG. 17.1. Temperature as a function of time in a system absorbing energy at a constant rate while undergoing phase transitions.

As we watch our ice container, we notice that, during the period of time while the temperature remains constant at 0°C, we have neither pure ice nor pure water but a mixture of both "phases." As soon as the last piece of ice has melted, our temperature begins to rise again but the slope is not so steep as before, indicating a rise in specific heat. At 100°C the temperature again stops rising, this time for a longer period of time, and finally, as soon as the last drop of water has been transformed into steam, the temperature of the remaining steam begins to rise. During the second period of constant temperature, we notice that we don't have pure water but rather a mixture of steam bubbles and water (*i.e.*, our water is boiling), just as we previously had a mixture of ice and water during the period of constant temperature.

If in our imaginary experiment the heat generator delivers 1 cal each minute, and if the original mass of ice is 1 gm, we see from Fig. 17.1 that 80 cal must be delivered during the first period of constant temperature at 0°C and 540 cal during the second period at 100°C. What happens to this heat?

The idea of indestructible energy offers a convenient interpretation of these observations: We invent a name for a new form of energy; we call it the *latent heat* **of fusion** and the *latent heat* **of vaporization.** *The former, amounting to* 80 *cal/gm in our case, is the amount of heat absorbed in the process of conversion of ice of* 0°*C into water of* 0°*C; the latter,* 540 *cal, is the amount of energy required to convert* 1 *gm of water at* 100°*C into* 1 *gm of steam at the same temperature.* This statement, of course, does not explain anything; it does not really tell how the absorbed energy was stored; it merely asserts that it is not lost and can be recovered by condensing our steam into water or by allowing the water to freeze to ice.

One defines the *heat of fusion* **as the amount of heat in calories absorbed in the process of melting of 1 gm of a given solid at constant temperature.**

The *heat of vaporization* **is defined as the heat absorbed in the process of transition of 1 gm of a liquid into the gaseous state at constant temperature.**

Conversely, the amount of heat liberated when 1 gm of the gas condenses (heat of condensation) is equal to the heat of evaporation and the heat liberated when 1 gm of the liquid solidifies is equal to the heat of fusion.

How could we demonstrate convincingly that heat is evolved in the formation of ice from water? This can be done simply by undercooling water, say to −10°C or so.

When we then shake it vigorously, the temperature suddenly rises to 0°C while at the same time ice crystals appear in the water. Obviously, heat was needed to produce this temperature rise. The observed rise in temperature can be accounted for quantitatively by the heat evolved in the formation of the observed amount of ice crystals.

We shall ascribe to a gram of water at 0°C a higher **internal energy** than to a gram of ice at 0°C and to a gram of water vapor a higher internal energy than to a gram of water at the same temperature. Whenever water vapor is condensed or water is frozen to ice, the difference in internal energy between the two phases is liberated as heat of condensation or heat of freezing, respectively.

The psychologically unsatisfactory feature of the energetic "explanation" of the phenomena of fusion and evaporation lies in the fact that it offers no mechanical picture of how the absorbed energy is stored. We shall intepret later the mechanism of these phenomena on the basis of the kinetic theory of matter.

Liquefaction of Gases

Instead of effecting a change of phase by heating, as has been done in the imaginary experiment of the preceding paragraph, one can produce changes of phase by means of pressure changes. Let us consider the following experiment. A gas, say carbon dioxide, is placed in a cylinder where it is compressed at *constant* room temperature of 21.5°C by moving the piston to the left (Fig. 17.2). We see from the 21.5° curve that the gas pressure increases up to a point *b*. From then on a reduction in volume does not cause an increase in pressure until a point *a* is reached. From then on the content of the cylinder behaves as if it were nearly incompressible. The slightest diminution in volume causes a tremendous increase in pressure, so that the curve rises nearly vertically to the left of *a*. The explanation of this behavior follows from the observation of the state of the cylinder content. As soon as the pressure corresponding to point *b* is reached, the gas begins to condense and continued reduction in volume leads to condensation of more gas rather than to an increase in pressure. Since a considerable amount of heat of condensation is given off in this process, it is necessary to withdraw this heat as fast as it is produced by cooling the cylinder if the process is to be conducted **isothermally,** *i.e.*, at constant temperature.

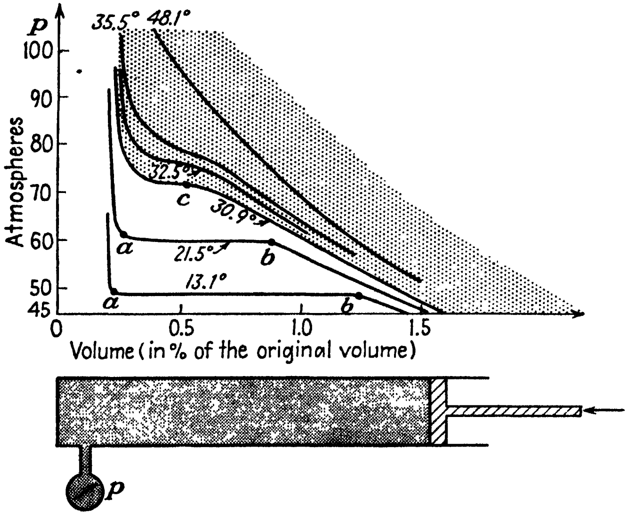


FIG. 17.2. Isothermal compression of carbon dioxide at different temperatures (given in degrees centigrade).

When the point *a* is reached, all the gas has been transformed into the liquid phase. Any attempt to reduce the volume beyond this point is an attempt to compress a liquid, the compressibility of which is extremely low.

The constant pressure, which is determined by the height of the plateau *a* — *b*, is called the **vapor pressure** of the liquid carbon dioxide.

The gaseous phase, which is in equilibrium with the condensed droplets of carbon dioxide, is referred to as the **saturated vapor** of CO_2. If the volume and temperature remain constant, the vapor and liquid phases coexist in the cylinder without transformation of one into the other; they are said to be in *equilibrium.* As long as there is any liquid CO_2 in the cylinder, an isothermal increase or decrease in volume will merely cause condensation of gas or evaporation of the liquid but no change in pressure. **A gas which is in equilibrium with its liquid phase is called a saturated vapor.** *The pressure corresponding to the plateau a — b is the saturated vapor pressure.*

As we repeat the above experiment at various temperatures, we notice that for higher temperatures of isothermal compressions (compare curves for 13.1 and 21.5°) the points a and b move closer together. Finally, at a certain temperature t_c (in this case 30.9°) both points a and b coincide at one point c. This is called the **critical point.** The corresponding temperature and pressure are referred to as the **critical temperature** and the **critical pressure.**

To understand the meaning of this observation let us recall the significance of the three portions of the curves in Fig. 17.2. In the portion to the right of point b the substance exists only in the gaseous phase. To the left of point a it is only in the liquid phase. The horizontal region a to b is the region of condensation, *i.e.*, of transition from the gaseous to the liquid state at constant pressure. At the critical temperature this horizontal region shrivels into one point c at which the tangent is horizontal. At this point liquid and gas can coexist at one definite pressure p_c and temperature T_c. At higher temperatures (*e.g.*, 32.5°) the curve has no horizontal section or even tangent at any point. There is no region of transition from gas to liquid. **Above the critical temperature it is impossible to liquefy a gas at any pressure.**

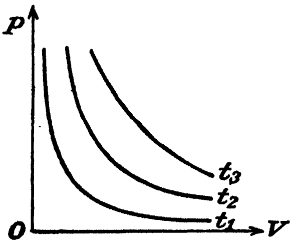


FIG. 17.3. Isothermal compression of an "ideal" gas at different temperatures (t_1, t_2, and t_3).

The region of our diagram to the right of the critical isotherm (30.9°C) is shaded. The isotherms for temperatures below T_c lie in the unshaded area. *A gas at a temperature below the critical T_c is often referred to as a vapor.*

The isotherms (graphs of p versus v at t = constant) show a diminishing inflection as the temperature increases. At 48.1°C the CO_2 isotherm in Fig. 17.2 is quite smooth. At sufficiently high temperature the isotherms become hyperbolas determined by the equation pv = constant (Boyle's law, see Chap. 18), such as shown in Fig. 17.3.

All gases show the pattern of behavior described above. *Any gas can be liquefied by pressure if it is cooled below the critical temperature.* For some gases the critical temperature is quite low as can be seen in the following table:

Substance	Critical temperature, °C	Critical pressure, atm
Water	374	218
Ammonia	132	112
Carbon dioxide	30.9	73.0
Oxygen	−119	49.7
Hydrogen	−240	12.8
Helium	−268	2.26

Effects of Pressure on Boiling and Melting

Water evaporates at room temperature and cools in this process. The rate of evaporation increases with temperature. At 100°C every gram of water absorbs 540 cal (the largest heat of evaporation of all known substances) when it changes into vapor of the same temperature. At lower temperatures the heat of evaporation is somewhat larger (about 600 cal/gm at 0°C). Cooling by evaporation has found many uses. For example, in hot countries drinking water is kept in porous bags which allow some water to be "sweated" out and evaporated. Each gram of the evaporating water is accompanied by the escape of about 600 cal. This heat is withdrawn from the water in the bag, as a result of which its temperature drops.

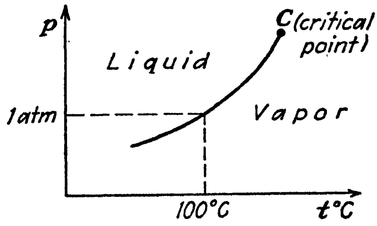


FIG. 17.4. Vapor pressure of a liquid as a function of temperature. (Why does the curve cease at point C?)

In general, any liquid evaporates in an atmosphere in which the pressure of its vapor is less than the saturation pressure corresponding to the given temperature. As the temperature increases, the vapor pressure of the liquid rises, as shown in the vapor-pressure curve for water in Fig. 17.4, and the rate of evaporation is increased. At a certain temperature a sudden change takes place in the character of the process. The evaporation ceases to be confined to the surface, and vapor bubbles appear in the interior of the liquid and rise to the surface. The liquid is said to be **boiling.** The temperature at which this occurs is called the **boiling point.** The boiling point depends on the atmospheric pressure. **A liquid boils when its vapor pressure becomes equal to the atmospheric pressure.**

Thus it is possible to make water boil at room temperature and even at 0°C by placing it in a closed container which is connected to a pump and reducing the pressure in the container by pumping. In fact it is possible to make water boil and freeze at the same time. If we start with water at room temperature, the process of evaporation will soon cool it to the freezing point, or putting it differently, the 600 cal supplied to every gram of water which is transformed into vapor is withdrawn from the liquid, whose temperature drops as a result. When the freezing point is reached, the required heat of evaporation cannot be provided by a further temperature drop of the liquid. Instead, the formation of ice connected with a liberation of 80 cal/gm provides the required heat of evaporation. Since the heat of evaporation (about 600 cal/gm) is much higher than the heat of fusion (80 cal/gm), much more water is frozen than evaporated in this process. Thus the water is made to freeze by boiling, and the temperature and pressure at which this occurs are $t = 0.0072°\text{C}$ and $p = 0.46$ cm Hg. Why not at 0°C? As we shall see below, the melting point of ice or, which is the same, the freezing point of water depends on the external pressure. It is 0°C at the atmospheric pressure of 76 cm Hg and is higher, namely, 0.0072°C at the vapor pressure of 0.46 cm Hg.†

As can be seen from Fig. 17.4, water can be made to boil at temperatures far above 100°C if it is subjected to pressures above 1 atmosphere, *i.e.*, 76 cm Hg. It can be also heated to temperatures above 100°C without boiling if the water surface is subjected to a pressure exceeding the vapor pressure corresponding to the given temperature according to the graph of Fig. 17.4. This possibility is utilized in pressure cookers, where food is heated in containers in which the vapor pressure of water is allowed to rise high above the atmospheric value, thus raising the temperature of the contents above 100°C. The same principle is also utilized in sterilizers of medical instruments, where under a gauge pressure of 25 lb/in.^2 medical instruments are exposed to steam of 130°C.

The melting of a solid which is in equilibrium with its liquid is affected by pressure to a lesser extent than the process of boiling. **In general, the melting temperature increases with increasing pressure for substances which contract on freezing and decreases with increasing pressure for substances which expand on freezing.**

---

† The condition of $p = 0.46$ cm Hg and $t = 0.0072°\text{C}$ is known as the *triple point* of water. At this temperature the pressure of the vapor, water, and ice are the same and the phases coexist in equilibrium. The equilibrium is disturbed when outside pressure is reduced below 0.46 cm Hg.

The former group of substances is more extensive. Water belongs to the latter group. A fusion curve can be drawn for water similar to the vapor-pressure curve. Figure 17.5 shows such a fusion curve representing the temperature of the melting (or freezing) point as a function of pressure. We see from this diagram that there is a slight drop in the melting temperature for a large increase in pressure. This fact plays an important role in some applications such as ice skating. Under the high pressure which is exerted upon the ice by the blade of a skate, the melting point may be lowered enough to make the ice melt at the prevailing low temperature. The thin film of water thus formed between the blade of the skate and the groove in the ice makes smooth sliding on the ice possible. As soon as the water thus formed escapes from the region of high pressure under the skate, it freezes again ("regelation"). The tendency of ice to melt under high pressure makes large masses of ice and snow deformable. This explains the formation of *glaciers*, large rivers of ice and snow which "flow" down "glacial" mountain valleys and erode them in a typical fashion. The flowing of these solid masses is due to the fact that the pressure of the enormous weight melts the solid mass locally, enabling it to adapt itself to the shape of an external constraint such as a mountain valley.

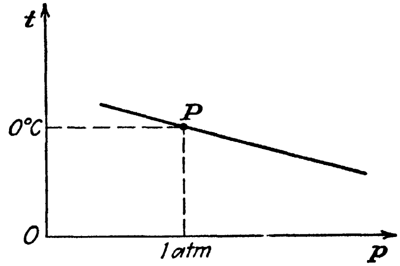


FIG. 17.5. Temperature of melting point as function of pressure for water.

Effects of Dissolved Substances on Boiling and Melting

Measurements of vapor pressure show that it is decreased for a liquid after certain substances have been dissolved in it. This implies that its boiling point should be elevated. One has to heat the liquid now to a higher temperature to reach a vapor pressure equal to atmospheric pressure.

The melting (or freezing) point is also affected by dissolved matter, being lowered in proportion to the amount of substance dissolved. The freezing point of a concentrated table-salt solution may be as low as −20°C.

It has been found that both the elevation of the boiling point as well as the lowering of the melting point are proportional to the number of gram molecular weights of the substance dissolved per unit volume, being quite independent of its chemical nature. This has enabled chemists to determine the molecular weights of compounds by measuring the elevation of the boiling point or the freezing-point depression which they produce in solution.

The lowering of the melting point may be used as follows in preparing a "freezing" mixture. Suppose we have a mixture of ice and pure water in phase equilibrium at 0°C. Let us assume that we now add salt to our mixture in such a concentration that the freezing point of the resulting solution would be at $-15°C$. This means that $-15°C$ is the temperature at which ice would be in equilibrium with the resulting salt solution. We see that the equilibrium of our ice-water mixture has been disturbed by the addition of salt. Ice begins to melt, and the heat of fusion required for the conversion of each gram of ice into 1 gm of water is withdrawn from the environment of the ice, *i.e.*, from the salt solution, whose temperature begins to drop and continues dropping until all the ice is dissolved. If we neglect the fact that the melting of ice dilutes the salt solution, we shall expect the temperature of $-15°C$ to be reached if there is enough ice in the mixture. When this temperature is reached, the ice stops melting and remains in equilibrium with the solution. By this method temperatures as low as $-20°C$ can be reached with ice or snow and salt.

Examples. 1. *Temperature Regulation of the Human Body.* The cooling effect of evaporation is of the utmost importance in human temperature regulation. In summer we manage to maintain our body temperature below the temperature of the environment owing to copious evaporation of perspiration. Exhalation of water vapor is an additional source of heat losses. (It is the primary factor in the heat regulation of dogs in hot weather.)

The motion of air is very important in the regulation of body temperature. In still air the evaporating vapor soon saturates the air surrounding the skin with water vapor, and evaporation is thereby impeded; a wind, however, constantly replaces the saturated air by dry air, thus maintaining a rapid rate of evaporation of sweat. It was observed that heatstroke is usually preceded by the exhaustion of the sweating mechanism.

In some countries, bakers' assistants regularly enter ovens at temperatures over 125°C well above the boiling point of water, and in some experiments people reportedly withstood temperatures above 225°C. In all these cases the hot air was quite dry, which facilitated rapid evaporation of perspiration. Prolonged exposure to air saturated with water vapor would be lethal at such temperatures.

Metalworkers have observed that they can dip a hand briefly into very hot molten metal without ill effect. This is due to the fact that almost instantly the hand is surrounded by a glove of vapor which is generated by evaporation of sweat on contact with the hot metal. Gases are very poor conductors of heat, and even a thin layer offers adequate heat insulation in this case.

2. *Local Anesthesia by Freezing.* Rapid evaporation of ethyl chloride, which is sprayed on gum tissues, is used by dentists to cause local anesthesia by refrigeration of tissues.

3. *Protection of Fruit from Freezing.* The freezing point of animal tissues, fruits, and vegetables is below 0°C due to the dissolved matter. Farmers make use of this by placing large containers of water in a basement in which fruits are stored. In cold weather the presence of water in the room prevents the temperature from sinking

below 0°C until all the water is frozen and thus protects the fruit, which could not stand much lower temperatures.

4. *Geological Processes.* Evaporation and freezing play an important role in the shaping of the crust of the earth. Rain water filling crevices in stones may freeze. Expansion of the freezing water cracks the stones. This process is effective in wearing down mountains in the course of geological ages.

Water which comes down as rain flows down river valleys, carrying the stone debris produced by the above-mentioned process of "weathering" into the oceans. The river water and ocean water evaporate and provide the material for the formation of clouds. The clouds cause again precipitation of rain, so that this "water cycle" continues without end, wearing down the mountains and elevating the bottoms of the rivers and ocean shores.

TRANSFER OF HEAT

Conduction of Heat

When we step on a cold metal plate and then on a wooden floor at the same temperature, the wooden floor feels less cold. Our feet are not cooled so rapidly on the wooden floor as on the metal plate; *i.e.*, heat is conducted away from our skin more rapidly when it is in contact with the metal than when it touches the wood. We express this by saying that the metal is a better conductor of heat than wood.

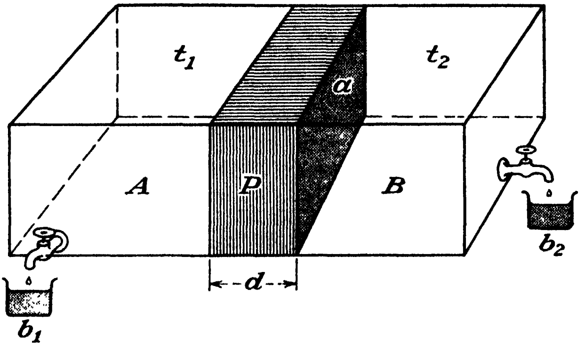


Fig. 17.6.

In order to form a clear quantitative idea of what is meant by conduction of heat, let us consider the following imaginary experiment (Fig. 17.6). A plate P of cross-sectional area a square centimeters and thickness of d centimeters is placed between two containers A and B which are maintained at constant temperature. The container A is in communication with a source of steam of 100°C, whereas the container B is filled with ice at 0°C. Thus, a temperature difference of $t_1 - t_2 = 100°C$ is maintained across the plate. We assume that the containers A and B are so well insulated that heat exchange can occur only across P.

After establishment of a steady state (*i.e.*, a constant temperature distribution across P), we shall observe steam to condense at a constant rate in container A. The rate of condensation can be established by catching the condensation water in the beaker b_1. As water condenses, heat is evolved at the rate of 540 cal/gm. What happens to this heat liberated in container A? We observe that ice is melting simultaneously in container B. The amount of water collected in beaker b_2 indicates that the number of calories absorbed in the melting of the ice in vessel B is equal to the heat liberated in A. It looks as if the heat generated in A were "flowing" through the plate P to the vessel B.

We can now investigate how the rate of heat flow, or heat transfer, depends on such factors as area of the plate, its thickness, the material, and the temperature difference maintained across it. Such investigations led to the following relationship:

$$\boxed{H = Ka\frac{t_1 - t_2}{d}} \tag{17.1}$$

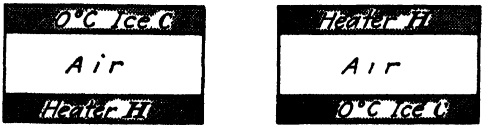
In this equation H is the number of calories which flow (in the sense of the above explanation) per second across the plate P, a is the area of the plate cross section, $t_1 - t_2$ is the temperature drop across P, d is the thickness of the plate, and K, the **coefficient of heat conduction,** is a constant characteristic of the material used. A study of different materials reveals wide variation in the values of K as shown in the following table:

Material	K, *(cal)(cm)/(sec)(cm²)(deg C)*
Silver	0.97
Copper	0.92
Aluminum	0.49
Brass	0.26
Steel	0.12
Ice	0.004
Glass	0.002
Cork	0.0001
Felt	0.0001
Air	0.000057
Hydrogen	0.00033

Materials with a high value of K are called good heat conductors; those with a low value of K, good insulators. All metals are good conductors in comparison with nonmetals.

Convection of Heat

Suppose we perform the following experiment (Fig. 17.7). The bottom of a box contains a heating element H, and the top an ice compartment C. When the heater is turned off, the air in the box is at 0°C due to contact with the ice. When the heater is turned on, the ice begins to melt, and the rate of formation of water in compartment C is a measure of the rate of heat transfer from H to C. (We neglect heat exchange with the environment.)



(a) (b)

FIG. 17.7.

We find that the rate of heat transfer across the air gap depends on the position of the device shown in the figure. It is faster when the ice is above the heater (Fig. 17.7) than when the ice is below it. This effect is obviously due to gravity. How can it be explained? To explain it let us consider the following experiment (Fig. 17.8): A large beaker containing some particles (whose density is nearly equal to that of water) is heated by a bunsen burner at the center of its bottom. The water in the beaker begins to move in a regular fashion as indicated by the motion of the suspended particles. A circulation develops as shown in Fig. 17.8. Why? We know that water expands when heated. Hence its density diminishes. A blob of hot water behaves in colder water similarly to a blob of oil, which is less dense than water—it rises due to buoyancy. But the hot water which rises upward from the heated point P does not leave an empty space there. Colder water rushes in from the sides to take its place; and so the hot water rises, displacing the water ahead of it and creating space for water to rush in behind it, thus giving rise to the circulation shown in the figure.

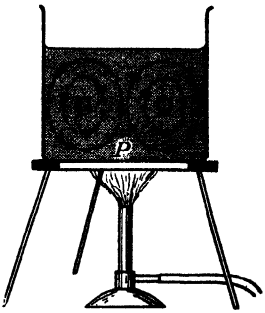


FIG. 17.8. Visualization of convection currents in water.

Another way of looking at it is as follows: Imagine a cylindrical column of water over the point P. When the water temperature is the same throughout the beaker, this column is in equilibrium with the surrounding fluid. We can consider it as a leg of a peculiar "U tube" which does not look like a U. The cylinder is the "inner leg," and the water surrounding the cylinder is the "outer leg." Now as the inner cylinder is filled with hot fluid of lesser density, its surface should rise above the surface of the outer leg for equilibrium. This would give rise to a small "hill of water" at the center of the water surface of the beaker, which we know cannot persist. The water will rather flow outward from the center near the surface in the opposite manner to the way the fluid moves toward the point P near the beaker bottom. Fluid currents set up in this manner are called **convection currents.**

Phenomena of this kind play an important role in nature as a cause of winds. For example, the atmosphere over an island which is heated by the radiation of the sun behaves similarly to the water over the point P in our beaker of Fig. 17.8. Radiant energy coming from the sun is absorbed in a very thin layer of the ground surface and heats the land high above the temperature of the surrounding ocean, which remains rather constant. The air which is in contact with the soil is heated and rises over the island, setting up convection currents similar to the currents seen in the beaker. As a result, during the daytime air moves toward the island near the surface, giving rise to the "sea breeze." (At night the motion of the air is reversed. Why?)

On a small scale we can observe convection currents in air over a radiator. These currents usually carry a good deal of dust upward and leave an unwelcome record in the form of a black mark on the ceiling above the radiator.

We are now in a position to interpret the results of the experiment illustrated in Fig. 17.7. In the case illustrated in Fig. 17.7*a*, convection currents of warm air rise from the heater and give off heat to the ice compartment. **Heat is transferred by convection.** In the case shown in Fig. 17.7*b*, on the other hand, the hot air remains in the top layer next to the heater, so that heat transfer can occur only by conduction and radiation (see page 252). Conduction is not very effective because of the low value of K for air (see table, page 250). Radiative heat transfer is also considerably less effective than convective transfer in this case.

In using insulating materials, such as rock wool, fur, or woolen blankets, one actually makes use of air as an insulating material. The wool has the function merely of preventing the formation of convection currents.

Heat Radiation

The modes of heat transfer considered so far required a "bridge" of matter (conduction) or material "messengers" (convection) in order to transmit heat from a hot body to a cold one. But there is a familiar mode of transmission of heat in which energy is transferred through empty space. The sun is a hot body. The energy which comes to us from the sun through the interplanetary vacuum is the source of the energy of the plants, animals, and our industrial fuels, such as oil and coal, which are of plant origin. This process of heat transfer without intervention of a material medium is called **heat radiation.** We can feel heat radiation if we hold our hand near the side of a radiator or when we approach our cheek accidentally to the cheek of another person.

How does the amount of heat transferred depend on the temperatures of the interacting bodies? On what other factors besides temperature does it depend? Stefan induced from experimental data that the energy radiated per second by each unit area of a body at the absolute temperature T is given by the expression

$$\boxed{R = e\sigma T^4} \tag{17.2}$$

if the body does not receive heat by radiation from other bodies. This equation is known as **Stefan's law.** The total rate of energy emission by the body is equal to RA, where A is the surface area of the radiating body.

One of the constants of the equation (17.2), $\sigma = 5.67 \times 10^{-5}$ erg/sec cm^2 deg^4 is independent of the properties of the radiating surface, whereas e, the emissivity of the surface, depends on the nature of the radiating surface. Its value lies between 0 and 1. A surface of maximum emissivity, $e = 1$, is referred to as a black surface or **black body.** For a nonblack body, e is defined as the ratio of the rate at which the given surface emits energy to the rate at which energy is emitted by a black body. For example, for tungsten at about 2700°C $e = 0.35$.

If the radiating body of absolute temperature T_1 is placed in an enclosure whose temperature is $T_2°$ absolute, the rate at which it loses energy is given by the equation

$$\boxed{R = e\sigma(T_1^4 - T_2^4)} \tag{17.2a}$$

Imagine now a room whose walls, ceiling, and floor are kept at the constant temperature of T_R°K. Suppose we suspend in this room a small sphere of the initial temperature T_0. If $T_0 < T_R$, the sphere will gain energy by absorbing the radiation which issues from the hotter walls and strikes its surface at a faster rate than it loses energy by radiation. Eventually a temperature equilibrium is reached when the temperature of the sphere becomes equal to the wall temperature. At this stage the energy which "streams" in a second from the walls toward the sphere is equal to the energy which returns each second from the sphere to the walls. The latter is composed of two components: (1) reflected energy, (2) energy emitted according to Stefan's law.

Suppose now that we polish the surface of the body without disturbing its temperature by polishing. Would this disturb the normal equilibrium? If we changed nothing besides the reflectivity of the surface, more of the radiant energy striking the sphere should be reflected; hence, the temperature of the sphere should drop. We know from experience, however, that a change in the reflectivity of the surface does not affect the temperature of the body under such conditions; hence we must assume that another property of this surface must be changed at the same time as its reflectivity is changed. Namely, we must assume that *the emissivity of a surface decreases as its reflectivity increases* in such a way that the flow of energy from the sphere toward the walls does not change under the conditions of the above experiment. By similar reasoning we can predict that a diminution of the reflectivity of the surface should increase its emissivity. These predictions are borne out by experience: *Good reflectors are poor emitters, and poorly reflecting surfaces have a high emissivity.*

For example, if we take two identical tin cans, fill them with boiling water, and cover the surface of one of the cans with lampblack or some other material which will make the surface a poor reflector, we notice that the poorly reflecting tin can cools off more rapidly. This is evidently due to more rapid radiation losses.

Some people make the mistake of covering the shiny metal pipes of a hot-air furnace with a thin layer of asbestos in order to diminish heat losses. Actually this only diminishes losses due to conduction, which are less significant in this case, but it increases radiation losses, since the asbestos layer is a poor reflector and, hence, a good radiator.

The dewar flask, or *thermos bottle,* is a good example which involves considerations of all types of heat transfer. Its construction is shown in Fig. 17.9. It is a double-walled hollow glass vessel. The substance which is to be thermally isolated is placed inside the space A. The interior B is evacuated in order to prevent heat transfer by convection. Heat conduction through the glass is poor because of the long path through a thin glass wall. In order to diminish radiative heat losses the inner wall is coated with a silver mirror (indicated by dotted line). Thus the walls of the flask are good reflectors, and hence little heat is lost by radiation if the content of A is hot. If the content of the flask is colder than the surroundings, the flask absorbs little heat by radiation from the surroundings, most of it being reflected.

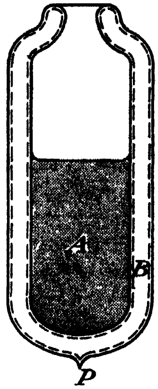


FIG. 17.9. Scheme of a thermos bottle.

Snow and ice can be made to melt rapidly by coating the surface with a good absorber of heat radiation. It has been suggested, for instance, to blaze ship lanes through arctic ice by spraying carpets of coal dust on the surface of the ice fields.

Spraying coal dust over soil increases the absorption of solar heat radiation and thus increases the rate of plant growth by warming the soil.

Cooling by radiation can also be easily demonstrated. If we place a pan of water on a poorly conducting base, such as a bundle of straw, on a clear night, the water will freeze even though the air temperature may be +2°C! How do we know it is not cooling by evaporation? This effect does not occur if we place the pan under a tree or if we perform the experiment on a cloudy night when the trees and clouds radiate energy into the pan.

THE SECOND LAW OF THERMODYNAMICS

There is probably no law of physics which has been formulated in so many different ways as the second law of thermodynamics. In its simplest formulation it sounds like an elementary law of heat transfer: **Heat cannot flow spontaneously from lower to higher temperatures.** Another formulation, **a perpetual-motion machine of the second kind is impossible,** looks hardly related to the first, but the relationship will become apparent from the discussion below.

What is a perpetual-motion machine of the second kind? It is a highly desirable contraption. It could, for instance (if it existed), withdraw heat from sea water (which would cool off in the process of yielding energy) and convert this energy completely into work. Thus, the propeller of a ship could be run at the expense of the internal energy of sea water. The fruitlessness of all attempts actually to devise such a machine led to the belief that such an engine is impossible in principle.

Experience shows that the conversion of heat into mechanical energy must be connected with a transfer of a certain fraction of heat energy from a high-temperature reservoir to one at lower temperature through the working substance of an engine. Whereas a given amount of mechanical energy can be converted practically completely into heat, a given amount of heat Q_1 can be converted only partly into mechanical energy in a periodic engine. The conversion is possible only if a certain amount of heat Q_2 is rejected into a low-temperature reservoir. In the case of a heat engine, Q_2 is the heat rejected in the exhaust of the engine. The difference $Q_1 - Q_2$ between the heat input Q_1 and the rejected heat Q_2 is converted into mechanical energy. The efficiency of the engine is defined as

$$e = \frac{\text{output of mechanical energy}}{\text{input of heat energy}} = \frac{Q_1 - Q_2}{Q_1}$$

Carnot found theoretically that no engine can exceed the efficiency of an ideal heat engine utilizing two heat reservoirs kept at temperatures T_1 and T_2. He found for the efficiency of such an ideal engine the expression

$$e = \frac{Q_1 - Q_2}{Q_1} = \frac{T_1 - T_2}{T_1}$$

T_1 is the absolute temperature of the hot reservoir from which the heat input Q_1 is supplied. T_2 is the temperature of the cold reservoir into which the nonutilizable heat Q_2 is rejected.

The inverse process is also possible; namely, a certain amount of heat Q_2 can be withdrawn from a low-temperature (T_2) reservoir, and a larger amount of heat Q_1 can be transferred to a reservoir at higher temperature (T_1) under expenditure of work. This sort of device is a refrigerator. This process does not violate the second law of thermodynamics as expressed by our first formulation, since in this case heat does not flow spontaneously "up a temperature hill" but is rather transferred through the working substance of an engine with an expenditure of work. For the variegated applications of the second law of thermodynamics the student is referred to texts on physical chemistry.

The discovery of the second law of thermodynamics is an interesting example of the induction of a general law of nature through considerations which were originally confined to a specialized engineering problem.

QUESTIONS AND PROBLEMS

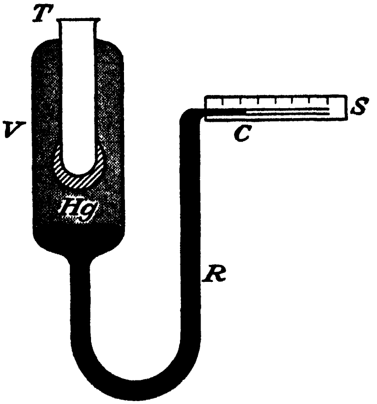
1. It was asserted in the text that stones may be broken by water getting into crevices and freezing there later on, the destructive action being due to the fact that ice has a larger specific volume than water. But does this explanation hold? Is it not correct to assume that ice will be simply squeezed out of the open crevice in the process of its formation without splitting the stone?

2. Is heat convection possible in a freely falling elevator?

3. Experiment shows that a candle goes out in a perfectly sealed falling box shortly after it is released. Explain why.

4. In a certain region of northern Canada the mean thickness of the layer of snow precipitated during a winter is about 50 cm. A layer of snow 1 cm thick when melted yields a layer of water 1 mm in thickness. How much heat is required to melt the snow covering an area of 1 m²?

5. The diagram shows Bunsen's ice calorimeter. A crust of ice *I* is frozen onto the bottom of the test-tube-like portion *T* of the apparatus. (How would you do it?) The part *V* is completely filled with water which rests on mercury. The mercury also fills the bent glass portion *R* which terminates in a capillary of 0.3 mm diameter. *S* is a scale. When a warm body is placed into *T*, some of the ice *I* melts. Since the density of ice is 0.9 gm/cm³ contraction takes place and is registered by movement of the mercury column past the scale *S*.



a. How large a motion of the mercury column will be obtained per calorie of heat absorbed by the ice?

b. What would be the maximum sensitivity of a calorimeter of this type that you consider practically achievable?

6. The density of sulfur at the melting point is $d_S = 2.05$ gm/cm³ for the solid phase and $d_L = 1.95$ gm/cm³ for the liquid phase at atmospheric pressure. How will the melting temperature be affected when sulfur is (*a*) subjected to a pressure of 10 atm? (*b*) placed in a vacuum?

7. A drink can be cooled by ice without being diluted by it if the ice is placed in a closed glass container which is submerged in the drink. An "ice cube" of this kind consists of a glass container weighing 20 gm and containing 200 gm of ice of −5°C. To what temperature will it cool ½ liter of mineral water of room temperature contained in a 100-gm glass flask (assume 20°C)? (Neglect heat exchange with the environment.)

8. In a refrigerating system which is based upon evaporation of liquid ammonia, how much ammonia must be evaporated in order to cool 20 kg of water from 20 to $-5°C$?

Liquid ammonia boils at 1 atm of pressure at a temperature of $-33°C$. The heat of evaporation is approximately equal to 320 cal/gm, and the specific heat of the ammonia gas is approximately equal to 0.52 cal/gm. Assume the thermal capacity of the machine to be 200 cal/deg.

9. 57 gm of steam of 100°C is blown from a boiler into a water-filled calorimeter in which the steam bubbles condense. What will be the final temperature if the copper calorimeter weighs 200 gm and contains 200 gm of water and 400 gm of ice?

10. A dentist usually heats his dental mirror before inserting it into a patient's mouth. Why?

11. If a barometer were filled with ether rather than mercury, how high would the ether column be at room temperature (say 20°C)? The density of ether is 0.74, and its vapor pressure at 20°C is approximately 44 cm Hg.

12. Why do we feel comfortable in a room at 70°F but find a bath in water of the same temperature cold?

13. Would it be equally efficient to install a radiator near the ceiling as near the floor? Explain.

14. CO_2 is a product of combustion of coal. It is also an excellent fire-extinguishing agent. Why does a coal fire not suffocate in its own combustion products?

15. How many calories of heat will escape per hour through a glass window 100 cm by 150 cm, 2 mm thick, if the outside temperature is $-10°C$ and the room temperature 20°C?

16. The rate at which energy is received on the surface of the earth by radiation from the sun is 2 cal/(min)(cm^2). How much solar power reaches the roof of a factory 100 by 100 m? Does it seem to you that solar energy could become an appreciable source of industrial power if it could be converted fully into useful work?

17. A tungsten filament of an electric lamp is heated to $t = 2430°C$. The length of the filament is 5 cm, and its diameter 0.2 mm. How much heat does the filament lose by radiation in 10 min in a room at 20°C? (Assume the emissivity of tungsten to be equal to 0.35.) How much would be radiated by a black body of equal area and temperature?

18. 10 cal/sec escapes through a hole of 2 cm^2 area in a furnace. What is the centigrade temperature of the furnace? (Treat the hole as a black body.)

19. A beaker contains 200 gm of ice in 1 liter of water. Salt is added until all the ice has melted. What is the final temperature of the solution in the beaker? (Neglect heat exchange with the beaker.)

★20. When a tiny ice crystal is dropped into a given test tube containing 20 cm^3 of undercooled water, 2 gm of ice is formed. Find the original temperature of the water.

21. What is the critical temperature of an ideal gas?

CHAPTER 18

THERMAL BEHAVIOR OF GASES

a. In what respects are the thermal properties of gases similar?
b. Can we cool an "ideal" gas by allowing it to expand into a vacuum?
c. Under what conditions will an ideal gas cool when expanding?
d. Why does a gas get hot when we compress it?
e. Why is the specific heat of a gas which can expand at constant pressure higher than when tested at constant volume?
f. How do we define "adiabatic" changes?
g. What must we do in order to (1) expand, (2) compress a gas isothermally?
h. Does the value of the constant in Boyle's law depend on temperature?
i. What assumptions regarding the compressibility of a gas do we make in setting up van der Waals' equation?
j. Why does a "real" gas cool when expanding in a vacuum?
k. Does the general gas law [equation (18.6)] hold only for isothermal processes, like Boyle's law, or is it equally applicable to adiabatic processes?
l. Is Poisson's law [equation (18.3*b*)] valid only at constant temperature?
m. What is osmosis?
n. How can we measure osmotic pressure?
o. Can work be performed by the process of osmosis?
p. Would you expect a temperature change when you dissolve a salt in water?
q. What is the difference between the "specific" and the "universal" gas constant?
r. Is the general gas law applicable to solutions as well as to gases?
s. What does Avogadro's law state?

Thus far, we have discussed the thermal properties of solids and fluids but have not considered the behavior of gases except for thermal expansion.

We have seen on that occasion that all so-called *permanent* gases (gases like oxygen, hydrogen, helium, etc., which can be liquefied only by great increase in pressure and reduction in temperature) have the same coefficient of thermal expansion and that this is equal to their temperature coefficient of pressure, namely, $\frac{1}{273}$ (or, more exactly, 1/273.2). This uniformity in behavior of gaseous matter was in striking contrast to the variety of coefficients of expansion characteristic of solid and liquid materials, which naturally leads us to the following interesting question: Does this uniformity in behavior of gases extend beyond the phenomena of thermal expansion and pressure variation? It turns out that a set of laws can be set up which apply to all gases of the aforementioned type regardless of their chemical identity. It seems natural to assume that this uniformity in the behavior of gases is not a coincidence but must have a common root in the peculiar nature of the gaseous state. Our attempts to explain the properties of gases will lead to a new method of physical explanation. It will be our first attempt to set up a hypothesis and to develop a theory. In order to prepare the factual material for such speculations some of the most important properties of gases will be reviewed in this chapter.

Energy Transformations

Every cyclist knows that his pump gets quite hot when he pumps up a tire. Compression heats the gas. In fact, the temperature elevation due to a sudden compression of a gas (for instance, compression in a cylinder, Fig. 18.1) may be made high enough to ignite a wick. (An outmoded type of cigarette lighter made use of this effect. The temperature elevation due to compression is also utilized in the diesel engine to ignite the fuel mixture in the cylinder.)

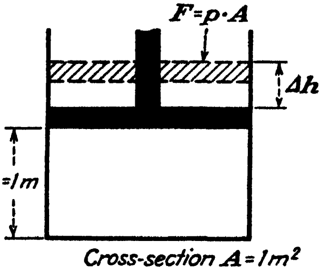


FIG. 18.1.

Although we are not yet in a position to suggest a mechanism through which compression generates heat, we can easily account for our observation by the law of conservation of energy. In compressing the gas we do work in advancing the piston against the forces due to gas pressure. This work is converted into heat, which elevates the temperature of the gas.

We can now ask the converse question: Could we lower the temperature of the gas by abstracting energy from it instead of adding it to the gas, as in the previous example? This can be done. Merely allow the gas in the cylinder to do work. For instance, place a weight on the piston, and allow the compressed gas in the cylinder to lift it through the distance Δh. The work done is $W\,\Delta h$. The energy has been provided by the gas. Whereas a gain in energy by the gas is manifested by a temperature rise, the present loss of energy will be revealed by a temperature drop of the gas. Instead of putting a weight on the piston, we may simply allow our gas to expand against the pressure of the atmosphere, which also provides a compressing force upon the piston. **Thus we can cool a gas by allowing it to do work.**

The Specific Heats of a Gas

We shall now be able to understand why a gas has, instead of a single definite specific heat, an infinite number of specific heats which lie between a minimum value (specific heat at constant volume) and a maximum value (specific heat at constant pressure).

Suppose our cylinder of Fig. 18.1 contains 1 gm of a gas. If we lock the piston so that the gas volume cannot change and supply enough heat to raise the temperature of the gas by 1°C, we shall increase the internal energy of our gas volume by c_v calories, *which is its specific heat at constant volume by definition.*

If, on the other hand, we leave the piston of Fig. 18.1 free to move, the internal gas pressure being equal to the atmospheric pressure, the gas will begin to expand as soon as we start heating it. When its temperature has risen by 1°C, it will occupy a slightly larger volume at the same pressure (equal to the atmospheric pressure). But this time we have to supply more heat, namely, c_p calories, since, in addition to raising the temperature of the gas, we have to supply the energy for the work which the expanding gas performs. *c_p is called the specific heat at constant pressure.* In expanding, our gas moves the piston through the distance Δh. Work done against the constant atmospheric pressure is $E = F\,\Delta h$, where $F = pA$ is the force exerted by the atmosphere upon the piston. Hence we can write

$$\boxed{\underline{E} = F\,\Delta h = (pA)\Delta h = p(A\,\Delta h) = \underline{p\,\Delta V}} \tag{18.1}$$

(The work done = pressure times volume increment.) Since the specific heat at constant pressure exceeds the specific heat at constant volume by the amount of work done in expansion, we can write

$$\underline{c_p - c_v = p\,\Delta V} \tag{18.2}$$

(where ΔV is the change in volume of 1 gm of gas due to a 1°C temperature change at constant pressure.)

It is easy to see that we can vary the specific heat of a gas by allowing it to expand only partly, *i.e.*, less than it would have at constant pressure and thus obtain any specific heat we wish between the values of c_p and c_v.

Meyer's Computation of the Mechanical Equivalent of Heat

It was the consideration of the difference between the specific heats of a gas at constant pressure and at constant volume which led Robert Meyer to his computation of the mechanical equivalent of heat. His reasoning ran about as follows:

1. Assume our gas volume to be at 0°C and atmospheric pressure to begin with. The excess energy (expressed in calories) which we have to supply to 1 gm of a gas (in addition to its specific heat at constant volume) in order to enable it to expand at constant pressure is equal to $c_p - c_v$. If instead of 1 gm we have m grams of gas, we shall have to supply $\underline{m(c_p - c_v)\text{ calories.}}$

Now suppose that the cross-sectional area of the cylinder in Fig. 18.1 is $A = 1\ \text{m}^2$ and its height $h = 1$ m. We know that the mass of a cubic meter of helium[†] at standard conditions[‡] is $m = 178.5$ gm. The specific heats of helium are also known from measurements: $c_p = 1.25$ cal/(gm)(deg C) and $c_v = 0.754$ cal/(gm)(deg C). Hence, the quantity of heat supplied for the expansion is

$$\Delta H = m(c_p - c_v) = 178.5(1.25 - 0.754)$$
$$\Delta H = \underline{88.54}\ \text{cal}$$

2. Now, how large is the work done by the expanding gas? $E = p\,\Delta V$. The normal atmospheric pressure is $p_0 = 1.013 \times 10^6$ dynes/cm². The change in volume is, from equation (16.2*b*),[§]

$$\Delta V = \frac{1}{273.2} V_0\,\Delta t = \frac{V_0}{273.2} \text{ since } \Delta t = 1°\text{C in our case}).$$

In our example, $V_0 = Ah = 1 \times 1 = 1\ \text{m}^3 = (100)^3\ \text{cm}^3 = 10^6\ \text{cm}^3$

hence, $$\Delta V = \frac{10^6}{273.2}\ \text{cm}^3$$

Thus,

$$p\,\Delta V = 1.013 \times 10^6 \times \frac{10^6}{273.2}\ \text{dyne}\cdot\text{cm} = \frac{1.013 \times 10^5}{273.2}\ \text{joules}$$

From this follows the **mechanical equivalent of heat** or the work equivalent of 1 cal as the ratio:

$$\boxed{J = \frac{p\,\Delta V}{m(c_p - c_v)} = \frac{1.013 \times 10^5/273.2}{88.54} = 4.185\ \text{joules/cal}}$$

Adiabatic Processes and Poisson's Law

We now understand why a gas in a container which insulates it thermally from its environment becomes hot when we compress it and cools off when we allow it to expand by reducing the force exerted on the piston. **Such volume changes during which a gas can neither receive heat from its environment nor lose heat to it are termed *adiabatic* expansion or compression.** (Nearly adiabatic changes occur when the heat conductivity of the container is poor and when the changes occur so fast that no appreciable heat transfer can take place during the considered time interval.)

[†] The gas considered here is not the same as in Meyer's original computation; he was using less accurate data and got a somewhat different value for *J*.

[‡] *I.e.*, at 0°C and atmospheric pressure.

[§] Instead of $\frac{1}{273}$, the more accurate value of $\frac{1}{273.2}$ has been used for β of helium.

Poisson found that the relation between pressure and volume in adiabatic changes is given by the expression

$$p = \frac{\text{constant}}{V^k} \tag{18.3a}$$

or

$$\underline{pV^k = \text{constant}} \quad \textbf{(Poisson's law)} \tag{18.3b}$$

where $k = c_p/c_v$ and where the value of the constant depends upon the gas and the amount of gas considered.

Isothermal Processes and Boyle's Law

Boyle found experimentally that a much simpler relation exists between the volume and pressure of a gas **if we cool the gas as we compress it or heat it while it expands in such a way that its temperature remains constant. Such changes are termed** *isothermal.*

For isothermal changes the pressure of a gas is simply inversely proportional to its volume:

$$p = \frac{C}{V} \tag{18.4a}$$

or

$$\boxed{pV = C} \quad \textbf{(Boyle's law)} \tag{18.4b}$$

The value of the constant C depends upon the gas and its amount and temperature. The shape of the curve representing V as a function of p is a "hyperbola" (see Fig. 17.3).

All gases obey this law closely at a sufficiently low pressure and sufficiently high temperature. Many gases (such as O_2, H_2, and He) follow it very closely far below room temperature and up to pressures of many atmospheres. But even these gases begin to deviate from it as they approach liquefaction.

We shall call a gas which obeys Boyle's law precisely and whose temperature coefficient of expansion and pressure is $\beta = 1/273.2$ an **ideal gas.** This is the pattern of behavior which all gases approach at low pressures and high temperatures.

Law of van der Waals

A more general law, which also describes the behavior of **real gases** at low temperatures and high pressures, was proposed by van der Waals:

$$\underline{(V - b)\left(p + \frac{a}{V^2}\right) = \text{constant}} \quad \textbf{(van der Waals' law)} \tag{18.5}$$

This equation has a wider range of validity than Boyle's law. It holds for the compression of a real as well as of an ideal gas at constant temperature. The value of the constant on the right side of equation (18.5) depends on temperature, the gas used, and the amount of gas considered. For ideal gases the constants b and a are zero and the equation goes over into the expression of Boyle's law [equation (18.4b)]. The reason for the replacement of Boyle's V by van der Waals' $(V - b)$ is as follows: We know that we cannot make the gas assume the volume zero by increasing the pressure indefinitely (contrary to the purely mathematical conclusion from $V = \text{constant}/p$). At a certain pressure the gas liquefies and the resulting liquid is scarcely compressible. Van der Waals' equation, by introducing $-b$, expresses the fact that, no matter how much we increase the pressure, we cannot reduce the volume below the value of b. This can be seen as follows: For very high values of p, $(V - b) = \text{constant}/[p + (a/V^2)]$ drops nearly to zero, so that V drops nearly to b.

The reason for the addition of a/V^2 to p is as follows: As a gas approaches condensation, it becomes more compressible than Boyle's law leads one to expect, as if some unknown additional pressure were aiding the compressing force. There is a good reason to assume such a "hidden" pressure. We have seen that the surface of a liquid tends to contract because of (intermolecular) forces which pull the particles of the liquid surface inward. The same arguments considered in connection with surface tension of liquids (Chap. 11, Appendix) hold also for a gas, the surface in this case being the boundary between the gas and the container. It can be shown theoretically and experimentally that the surface tension of a gas volume is inversely proportional to V^2 and, hence, is not very noticeable except under strong compression. *Owing to this inward pull exerted on the surface of the gas the applied pressure p appears increased by the amount a/V^2.*

Joule-Thomson Effect

We are now in a position to understand an important phenomenon, the Joule-Thomson effect, which enables us to reach low enough temperatures to liquefy even such gases as air or helium. We saw previously that we can cool a gas by allowing it to do work while expanding against an external pressure. Now, an expanding "real" gas should do work in overcoming the forces of surface tension which act in the same direction as an external force of compression. We should therefore expect a "real" gas to cool as it expands in a vacuum. For instance, if we were to fill a glass balloon with, say, carbon dioxide at room temperature and place it in an evacuated chamber, we should expect the temperature of the gas to drop after we break the glass balloon and thus allow the gas to expand, filling the evacuated chamber. Joule and Thomson actually succeeded in detecting a drop in the temperature of a freely expanding "real" gas, whereas the temperature of an "ideal" gas does not change as it expands in a vacuum.† The temperature change obtained with freely expanding **real** gases is known as the **Joule-Thomson effect.**

The Joule-Thomson effect is used extensively in refrigeration engineering. With the aid of this phenomenon it is possible to cool gases such as air, oxygen, hydrogen, or helium (after they have been precooled to a temperature at which they behave like a "real" gas) to the point of liquefaction.

Pathologists utilize the Joule-Thomson effect by allowing compressed CO_2 to escape through a valve, striking a specimen of human tissue. This process, which cools the gas, allows it to solidify the tissue by freezing it and makes it possible to cut the latter into fine slices with a microtome for microscopic observations. (The temperature drop is also partly due to the work done by the expanding CO_2 against the atmospheric pressure.)

The General Gas Law

We shall now summarize our knowledge of the relationship between the pressure, volume, and temperature of an "ideal" gas sample in one single expression, the general gas law, which we can derive as follows: We start from the expressions for

Boyle's law: $p_1V_1 = p_2V_2 = \text{constant}$ (*a*)
Thermal expansion: $V_t = V_0(1 + \beta t)$ (where $\beta = \frac{1}{273}$) (*b*)

Let us now assume, for simplicity, that we start out with m grams of a gas at standard conditions, that is, $p_0 = 1$ atmosphere and $t_0 = 0°\text{C}$. We shall designate the gas volume under these conditions by V_0.

First, suppose we heat our gas, keeping the pressure p_0 constant. We shall then have the following transition:

$$\begin{array}{lccc l} \text{Original state:} & V_0, & t_0, & p_0 & (1) \\ & \downarrow & \downarrow & \| & \\ \text{Final state:} & V_t, & t, & p_0 & (2) \end{array}$$

After heating, our final volume V_t will be related to the original volume V_0 through equation (*b*) $V_t = V_0(1 + \beta t)$.

Next, we shall allow the volume V_t of our gas and its pressure p_0 to assume any pair of values which are compatible with the condition that the temperature remain constant. In other words, we make the transition

---

† The work done against surface-tension forces is zero for an "ideal" gas, since $a = 0$ in the a/V^2 term which represents the inward surface pull per unit area in van der Waals' equation.

$$\begin{array}{lccc}
\text{From the original state:} & V_t, & t, & p_0 \\
 & \downarrow & \| & \downarrow \\
\text{To our final state:} & V, & t, & p
\end{array}$$

(2)

(3)

Since this is an isothermal process (t remaining constant), the final and original values of pressure and volume should satisfy Boyle's law:

$$pV = p_0 V_t \tag{4}$$

Now we substitute (b) in (4):

$$\underline{pV = p_0 V_0 (1 + \beta t)} \tag{5}$$

and we have derived the desired law. It is, however, convenient to simplify the equation by introducing absolute temperature instead of centigrade temperature. ($T = t + 273$, where T is the absolute temperature and t is in degrees centigrade.) Since $\beta = \frac{1}{273}$, we can write

$$pV = p_0 V_0 \left(1 + \frac{t}{273}\right) = p_0 V_0 \left(\frac{273}{273} + \frac{t}{273}\right) = p_0 V_0 \left(\frac{273 + t}{273}\right)$$
$$= p_0 V_0 \left(\frac{T}{273}\right) = p_0 V_0 \frac{T}{T_0} \tag{5a}$$

(where we replaced 273 by T_0, which stands for the absolute temperature of melting ice).

Designating by p_1, V_1, and T_1 or p_2, V_2, and T_2, etc., any triplet of values of p, V, and T which satisfy equation (5a), we can write

$$\frac{p_1 V_1}{T_1} = \frac{p_2 V_2}{T_2} = \boxed{\frac{pV}{T} = \frac{p_0 V_0}{T_0} = \text{constant}} \tag{18.6}$$

This is the **general gas law.**

In words: **The ratio of the product of pressure times volume divided by the absolute temperature remains constant for an ideal gas.**

We obtain Boyle's law as a special case of this law by setting $T_1 = T_2$.

The value of the constant depends on the nature of the gas and on the amount under consideration. The gas constant for a gram of any gas is simply equal to $p_0 v_0 / T_0$, where p_0 is the standard atmospheric pressure, T_0 is equal to 273° absolute, and v_0 the volume of a gram of this gas at standard conditions of pressure and temperature. (The so-called specific volume = 1/density.)

Analogies between the Behavior of Solutions and of Gases

The behavior of solutions is in many respects strikingly similar to the behavior of gases. For instance, if we pour a concentrated solution of copper sulfate into a glass cylinder and then carefully pour clear water over it, we can obtain a sharp demarcation line between the water and the colored solution. But in a few hours the demarcation line is blurred, and we see that the copper sulfate seems to have spread into the space occupied by clear water much in the same way a gas would have spread to fill a vacuum or to penetrate a space filled by another gas. The most conspicuous difference between this **diffusion** of a dissolved substance and the analogous behavior of a gas is the relative slowness of the former process.

Just as an expanding gas can do work, a solution can perform work when allowed to "expand" (*i.e.*, to become more diluted). We can, for instance, perform the following experiment (Fig. 18.2). C is a glass cylinder, closed off by membrane M made of a pig's bladder. A solution, say of table salt or sugar, is poured into this cylinder, which is then dipped into a beaker of pure water so that the level A of the liquid in the cylinder coincides initially with the water level A'.

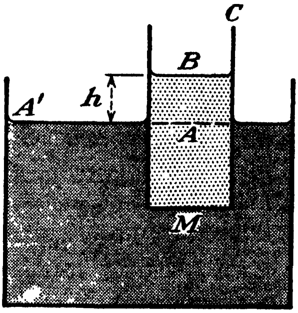


Fig. 18.2. Osmosis.

Very soon we notice that the level in the cylinder rises until it reaches some maximum height h. This case is reminiscent of the cylinder of a steam engine, with the dissolved substance playing the part of the expanding steam and the water level B acting as the piston upon which the solution exerts a pressure (the **osmotic pressure**) which lifts it. Water appears to pass rather freely through our membrane, whereas the dissolved substance scarcely penetrates it, as can be ascertained by tasting (or testing) the water in the outer beaker. A membrane which shows such a selectivity in passing the solvent but not the dissolved substance is called a **semipermeable membrane.**

The magnitude of the osmotic pressure is determined by the maximum height h to which the level B rises. (This is, of course, the osmotic pressure of the *final* concentration of the dissolved substance, which is less than the original concentration that prevailed while the inner liquid level was at A. How would you modify this arrangement in order to measure the *original* osmotic pressure?)

Diffusion and osmosis are perhaps the two most important physical processes in the maintenance of life, since it is through them that foodstuffs can be absorbed by an organism. Osmotic pressure is also (at least partly) responsible for the maintenance of rigidity ("turgor") of plants. The roots of a plant are provided with a semipermeable membrane. The environment of the roots is "hypotonic"; *i.e.*, its osmotic pressure is lower than that of the interior of the roots. Consequently, fluid is "pumped" by **osmosis** into the plant until the elastic back pressure equals the osmotic pressure. The action of this so-called *root pressure* is well demonstrated by the "bleeding" of a plant from the cut ends of a stem.

Osmosis is also important in secretion. The role of osmosis in this process is well demonstrated by the model of an artificial gland shown in Fig. 18.3. A glass cylinder C containing a salt solution is closed at end A by a semipermeable membrane and at end B by a permeable one which passes the salt as well as the water. The surrounding fresh water enters the cylinder through the membrane A by osmosis, and the solution is "secreted" from our cylindrical "gland" through the membrane B back into the tank as shown by the arrows. Our arrangement resembles the action of a circulating pump. The pumping work is done at the expense of the original internal energy of the concentrated solution in cylinder C, which becomes more dilute in the course of this process.

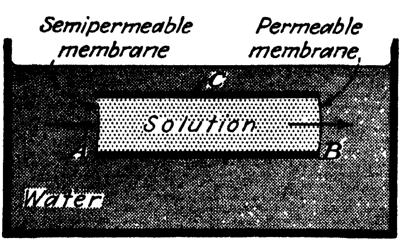


FIG. 18.3. Model of an artificial gland.

Osmotic volume changes can be very well observed with red blood corpuscles. When placed in a "hypertonic" solution (*i.e.*, a solution whose osmotic pressure is higher than that inside the cells), they shrink in size. In a hypotonic solution they swell, and if the solution is dilute enough, they will burst (hemolysis). In making rapid infusions into the blood stream it is important to use an "isotonic" solution of 0.9 per cent NaCl, the osmotic pressure of which is equal to that of the blood cells, in order to avoid hemolysis.

What interests us most about osmotic pressure are the following two physical features:

1. The osmotic pressure increases with temperature. The osmotic pressure at any temperature can be obtained from the pressure at 0°C through the relation $p = p_0(1 + \beta t)$, where the temperature coefficient of pressure is $\beta = \frac{1}{273}$. This is an amazing fact! This is precisely the law that holds for gases, and the value of β is not even different!

2. **The osmotic pressure of a solution is equal to the pressure which the dissolved substance would exert if it were present as a gas in the volume occupied by the solution.**

(Van't Hoff's law)

These are interesting but baffling empirical correlations between seemingly unrelated phenomena. Why do solutions resemble gases so closely? Why do all gases (at a sufficiently high temperature and low enough pressure) have the same coefficients of pressure and volume and follow the same general gas law? Why is the temperature coefficient of osmotic pressure equal to the temperature coefficient of gas pressure? Why does the gas pressure increase at all with temperature, and why is it inversely proportional to the volume at constant temperature? Why is the osmotic pressure of a solution proportional to the concentration of the solution?

If these facts have a common explanation, it is certainly not suggested by what we know so far. But we shall become acquainted presently with a few additional observations which throw new light on the entire picture and hint in which direction a unified explanation may lie.

Avogadro's Law and the Universal Gas Constant

Avogadro discovered that, if you compare various gases at the same temperature, the same pressure (say 0°C at atmospheric pressure), and the same volume, then you find that the ratio of the weights of the samples is equal to the ratio of the molecular weights found by chemists for these gases. In other words, if you choose $p_0 = 1$ **atmosphere,** $t_0 =$ **0°C and** $V_0 =$ **22.4 liters,** then, for this special value of V_0, **each sample will contain 1 gram molecular weight of the gas**; *i.e.*, you will have 2 gm of H_2, 32 gm of O_2, 28 gm of N_2, 44 gm of CO_2, etc.

At the same temperature and the same pressure equal volumes of different gases contain equal numbers of gram molecular weights.†

(Avogadro's law)

We can use this law to unify our general gas law so that one single constant would apply to all gases:

Let us write the general gas law for 1 gm of some gas. The volume of 1 gm is the "specific volume"; we shall designate it by v, whereas the volume of a mass of m grams will be symbolized by V. We write then for 1 gm

$$\frac{pv}{T} = \frac{p_0 v_0}{T_0} = C \tag{18.7}$$

where C **is the "specific" gas constant per gram** for a given gas. Multiplying (18.7) by m on both sides, we get (since $mv = V$)

$$\boxed{\frac{pV}{T} = \frac{p_0 V_0}{T_0} = mC} \tag{18.8}$$

---

† This is not the original formulation of this law by Avogadro, who used the terminology of the molecular theory to express it. We obtain Avogadro's original formulation by replacing in the above statement "gram molecular weights" by "molecules" (see Chap. 20).

Now, any mass m can be expressed in terms of the gram molecular weight of its substance. For instance, 32 gm of O_2 is 1 gm mole, 64 gm of O_2 is 2 gm moles, 8 gm O_2 is 0.25 gm mole. In general, if the mass is m and the molecular weight is M, we can set

$$\boxed{m = nM} \tag{18.9}$$

that is, m grams of the gas are n gram moles.

Substituting (18.9) in (18.8) we obtain

$$\boxed{\frac{pV}{T} = \frac{p_0V_0}{T_0} = nMC} \tag{18.10}$$

or

$$MC = \frac{p_0V_0}{T_0n} \tag{18.10a}$$

Now suppose you compare 1 gm mole of all gases. Then, $n = 1$ for all, and since p_0, V_0, and T_0 are fixed values, independent of the nature of the gas, we see that *for* 1 *gm mole of any gas*

$$\boxed{MC = \text{constant} = R} \tag{18.10b}$$

substituting (18.10*b*) in (18.10) we get

$$\boxed{\frac{pV}{T} = \frac{p_0V_0}{T_0} = nR} \tag{18.11}$$

R is the so-called **universal gas constant.** **$R = 8.31$ joules/(mole) (deg).** R is the same for *all* gases. n is the number of gram moles of substance contained in the gas sample. T is measured in degrees Kelvin, p in dynes per square centimeter, and V in cubic centimeters.

Numerous other examples can be given as further illustrations of how the gas laws can be unified (*i.e.*, one single equation made to hold without change of constants for all gases) if, instead of measuring masses in grams, we express them in terms of the number of gram molecular weights. For instance, the difference between the specific heats of a gas expressed *per mole* is the same for all (ideal) gases and is equal to the universal gas constant $C_p - C_v = R$ (see Prob. 1). Or the specific heat *per mole* at constant volume of all monoatomic gases is the same: $C'_v = \frac{3}{2}R$ (where R is the universal gas constant). The specific heat of diatomic gases is $C''_v = \frac{5}{2}R$. For most solid elements at room temperature the product of the specific heat and the atomic weight (*i.e.*, the specific heat per gram atom) is about the same, being equal approximately to 6 (law of Dulong and Petit).

Perhaps these relationships between the thermal gas laws (and the analogous laws for solutions) and the chemists' molecular weights point in the direction in which we should look for the explanation of the thermal properties of gases and of solutions and possibly even of the nature of heat. We shall examine in Chap. 19 the meaning of the chemical gram molecular weights and shall investigate whether or not hypotheses which the chemists have adopted for the explanation of chemical behavior might be helpful in interpreting and correlating physical phenomena, the presentation of which is simplified by the introduction of the molecular weights.

Example 1. Computation of the Universal Gas Constant

Solving equation (18.11) for R we get

$$R = \frac{1}{n}\frac{p_0V_0}{T_0}$$

If we consider $n = 1$ mole of any gas, the volume occupied at $T_0 = 273°\text{K}$ and $p_0 = 1$ atm $= 1{,}013{,}000$ dynes/cm^2 is $V_0 = 22.4$ liters $= 22.4 \times 10^3$ cm^3.

$$\text{Hence, } R = \frac{1.013 \times 10^6 \times (22.4) \times 10^3}{273} = 8.31 \times 10^7 \text{ ergs/(mole)(deg)}$$

$$= \underline{8.31 \text{ joules/(mole)(deg)}}$$

Example 2

A tank contains 10 liters of oxygen at 10 atm pressure at a temperature of 127°C. How many grams of oxygen does it contain?

Solution. $R = 8.31 \times 10^7$ ergs/(mole)(deg); $V = 10^4$ cm^3;
$p = 1.013 \times 10^7$ dynes/cm^2; $T = 127 + 273 = 400°\text{K}$

General gas law [equation (18.11)]:

$$\frac{pV}{T} = nR$$

$$n = \frac{pV}{T}\frac{1}{R} = \frac{1.013 \times 10^7(10^4)}{400} \times \frac{1}{8.31 \times 10^7} = 3.05 \text{ moles}$$

Now the atomic weight of oxygen is 16, and the molecular weight is $M = 32$. One gram mole of oxygen is 32 gm, and hence

$$n = 3.05 \text{ gm moles of oxygen}$$

corresponds to

$$\underline{m} = nM = 3.05 \times 32 = \underline{97.6 \text{ gm}}$$

Example 3

A 4-liter tank contains 39.0 gm of a gas at a pressure of 10 atm and a temperature of 400°K.

a. What gas is it (assuming it to be a chemical element)?
b. What is its density under the given conditions?
c. What is its density at standard conditions?

Solution. *a.* General gas law [equation (18.11)]: $pV/T = nR$. According to equation (18.9), $m = nM$; $n = m/M$, where M is the molecular weight of the gas and m the mass of the gas sample. Substituting (18.9) in (18.11) we get

$$\frac{pV}{T} = \frac{m}{M}R$$

and

$$M = \frac{mRT}{pV}$$

This equation expresses the molecular weight of our gas in terms of the given data. Given data: $m = 39.0$ gm; $R = 8.31 \times 10^7$ ergs/(mole)(deg); $T = 400°\text{K}$;

$$p = 1.013 \times 10^7 \text{ dynes/cm}^2; \qquad V = 4 \times 10^3 \text{ cm}^3$$

Thus

$$M = \frac{39.0(8.31 \times 10^7)400}{(1.013 \times 10^7)4{,}000} = 32$$

which is the molecular weight of *oxygen.*

b. Density is defined as $d = \frac{m}{V} = \frac{39.0}{4 \times 10^3} = 9.77 \times 10^{-3}$ gm/cm³

c. General gas law: $\frac{pV}{T} = \frac{p_0V_0}{T_0}$

Since $d = m/V$, that is, $V = m/d$, we get, substituting this value for V in the above equation,

$$\frac{p\cancel{V}}{Td} = \frac{p_0\cancel{V}}{T_0d_0}$$

Hence, $\underline{d_0} = \frac{p_0}{T_0}\frac{T}{p}d = \frac{1.013 \times 10^6}{273} \times \frac{400}{10(1.013 \times 10^6)} \times 9.77 \times 10^{-2}$

$$= \underline{1.43 \times 10^{-3} \text{ gm/cm}^3}$$

QUESTIONS AND PROBLEMS

★**1.** Prove, using equation (18.2), that $C_p - C_v = R$, where C_p and C_v are the specific heats at constant pressure and constant volume *per mole* of a gas and R the universal gas constant. (Hint: Express ΔV, the expansion at constant pressure per degree, in terms of the coefficient of expansion of a gas at constant pressure.)

2. The specific heat of oxygen at constant atmospheric pressure is

$$c_p = 0.218 \text{ cal/(gm) (deg C)}$$

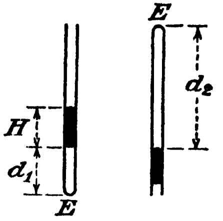
a. Find the specific heat at constant volume.

b. Find the ratio $k = c_p/c_v$. (The density of oxygen at standard conditions is $d_0 = 1.43 \times 10^{-3}$ gm/cm³.)

3. In an engine gas is compressed adiabatically to one-twentieth of its initial volume. If the initial pressure is 1 atm and the initial temperature is 200°C, find the pressure and temperature after compression. Do you have to use Poisson's law? Can you solve this problem using the general gas law? (Use $k = 1.4$.)

4. A bubble expands to three times its original volume in rising from the bottom to the surface of a lake. How deep is the lake? (Assume normal atmospheric pressure $p_a = 76$ cm Hg.)

5. A mercury column $H = 10$ cm long is placed in a capillary tube which is closed at end E. In the first position of the diagram the distance d_1 of the bottom of the mercury column from the bottom of the tube is $d_1 = 20$ cm at normal atmospheric pressure of $p_a = 76$ cm Hg. Find the distance d_2 for the inverted tube.



6. 100 cm^3 of hydrogen at 20°C is heated to 100°C and simultaneously compressed until the original pressure is doubled. Find the final volume.

7. A tire is rapidly pumped up to a gauge pressure of 32 lb/in^2. The air is thereby heated to an initial temperature of 40°C. Determine the pressure in the tire when the air cools off to 20°C. Neglect changes in volume of the tire.

8. How many grams of oxygen are contained in a 10-liter tank at a pressure of 10 atm and a temperature of 20°C?

9. What is the pressure in a 10-liter tank which contains 10 gm of hydrogen at a temperature of 25°C?

10. An oxygen tank containing 10 liters of oxygen at a pressure of 12 atm and a temperature of 20°C is heated in a fire accident to a temperature of 300°C. What is the pressure in the tank under these conditions?

11. Do you expect the Joule-Thomson effect to be more pronounced at high or at low temperatures? Why?

★12. A gas flows through a pipe of uniform cross section. Due to internal friction (viscosity) there is a pressure drop along the pipe so that the pressure near the west end is higher than near the east end.

a. Which way does the gas flow?

b. Is the linear velocity of flow the same at all points along the axis of the pipe?

c. Is the motion of an individual gas particle along the pipe (1) uniform? (2) accelerated? (3) decelerated?

13. Can you design a refrigerator making use of the Joule-Thomson effect?

14. In what ways is the behavior of solutions analogous to the behavior of gases?

15. Under what conditions will the specific heat of a gas be higher than at constant volume yet lower than at constant pressure?

16. Can you measure the osmotic pressure of a solution without disturbing it appreciably by the arrangement pictured in the text (Fig. 18.2)? If not, how would you modify it? How will the size of the membrane influence the height to which the liquid rises? How will the process of osmosis be affected if we double the area of the membrane?

17. Swellings due to edema may be reduced by application of wet epsom salt dressings. What physical explanation can you offer for this effect?

18. In a basal metabolism test a patient is found to be inhaling oxygen at the rate of 160 cm^3/min at a barometric pressure of 75.5 cm and room temperature of 28°C. Convert this volume to normal conditions of temperature and pressure.

19. Verify the law of Dulong and Petit for the cases of

	Pb	Li	Cu	Al
Atomic weight............	207.2	6.94	63.5	27.0
Specific heat..............	0.0314	0.94	0.095	0.214

> In the true philosophy . . . the causes of all natural effects are conceived as mechanical causes, which, in my opinion, must be done or all hope of ever understanding physics renounced.
>
> *Huygens*

PART B: Theoretical Approach in Physics Using Mechanical Model Theories

In Part A we have established the laws of motion of different forms of matter and the laws governing the thermal properties of matter. The two groups of laws appear to be entirely unrelated. What steps can be taken to achieve an understanding of both sets of laws? Our conception of understanding the laws of nature is as follows: We understand a group of natural laws if we see logical connections among them, for instance, if we can deduce them logically from a common set of assumptions or if we can deduce any law of the group from other laws of the same group. A set of assumptions from which laws can be derived is called a theory. Just as it is the purpose of a law to predict facts, the purpose of a theory is to predict laws. How does one choose the set of the assumptions of a theory? There are no fixed rules for that. The theorizing scientist has to use his imagination in addition to judgment. Once a set of assumptions is adopted from which known laws can be deduced, one proceeds to deduce yet unknown laws from them. At this point the experimentalist steps in and checks by experiment the correctness of the predictions of the theory. If all the predictions are verified, the theory is adopted as a "correct" theory. But this is a misleading term. The theory can be proved wrong at any moment, whenever we find a fact which is inconsistent with it. Such a finding leads either to abandonment or to a modification of the theory. Thus the experimentalist not only checks theories but also stimulates the formulation of new ones. And in what way does the theoretician influence the activity of the experimentalist? Without theories the search for new laws of nature would be a mere fumbling in the dark. Theories provide leads for experiments which are likely to result in the discovery of new laws. The attitude of modern physicists is that the sole justification of a theory is its success in predicting new laws rather than its plausibility. In the early days of the development of physics, however, the plausibility of a theory played an important role. The set of assumptions had to involve preferably some kind of a mechanical picture, so as to deduce the new, unfamiliar laws from the familiar laws of mechanics.

A logical bridge has been created between the laws governing the thermal behavior of matter and the laws of mechanics through the kinetic theory of gases. The dependence of certain thermal properties of gases on the molecular weight introduced the chemists' molecular conception of the structure of matter into physics. A mechanical model of a gas was designed in imagination, whose components (the molecules) moved according to Newton's laws of mechanics (Chaps. 19 and 20). This model enabled one to deduce all known gas laws plus some additional unsuspected laws.

Similarities between the definite, multiple proportions in which electricity is found associated with matter (Chap. 21) with the laws of chemical combination suggested the idea of "atoms of electricity," whose existence was eventually demonstrated by experiment (Chap. 22). In Chaps. 21 and 22 phenomena connected with transport of electricity through matter and vacuum are explained by the assumed migration of charged atoms of matter and atoms of electricity, which move under the influence of electric forces according to the laws of mechanics.

In Chap. 23 we encounter a new type of force between electric charges, which depends not only on their magnitude and distance but also on their relative velocity. The observed phenomena of electromagnetism (at first presented phenomenologically in Chaps. 23 and 24) are explained by the forces of interaction between the ions or electrons which generate a so-called *magnetic field* and ions or electrons which move through that field.

CHAPTER 19

KINETIC MOLECULAR THEORY: 1. QUALITATIVE PART

a. What is the objective of the kinetic molecular hypothesis?
b. What are the assumptions of the kinetic molecular hypothesis?
c. What is the distinction between a hypothesis and a theory?
d. What scheme or method does a scientist use in formulating a hypothesis?
e. Why do scientists try to get along with as few hypotheses as possible?
f. What criteria do we use in assessing the value of a hypothesis or theory?
g. How do we decide whether or not a theory is "correct"?
h. What evidence do we have for the assumption of granular (discontinuous) structure of matter?
i. What evidence led us to infer the existence of (1) atoms? (2) molecules?
j. What is (1) an atom? (2) a molecule?
k. On the basis of what evidence do we assume that molecules are in motion?
l. How do temperature changes affect molecular motion in an homogeneous body?
m. On what evidence did you base your answer to (*l*)?
n. When the kinetic energy of a moving object is converted into heat by impact, what happens to the original kinetic energy according to the kinetic theory of heat and matter?
o. How does the kinetic theory explain gas pressure?
p. How does the kinetic theory explain Boyle's law?
q. How does the kinetic theory explain heat conduction?
r. How does the kinetic molecular theory explain cooling by evaporation.
s. How does the kinetic molecular theory explain absolute zero?
t. What do you consider the most convincing evidence of molecular motion?
u. What is Brownian movement due to?

In considering transformations of different types of energy, the effects of heat on matter, and the common properties of gases, we encountered on the one hand a great variety of seemingly unrelated phenomena and on the other hand, as in the case of gases and solutions, some striking relationships which we could not explain on the basis of one single unifying principle.

Formulation of Hypotheses

The mere accumulation of facts is not the ultimate goal of science. The objective of the scientist is to systematize knowledge. He is constantly searching for causal relationships between known facts in the belief that all known facts are a consequence of a few basic laws of nature to which they can be related logically as well as to each other. If he fails to see a relationship between phenomenon *A* and phenomenon

C, he will expand his knowledge by observation in the hope of discovering a phenomenon B which would enable him to correlate A and C by correlating A with B and B with C. If observations do not lead him to the discovery of facts facilitating correlation, he resorts to his imagination to provide the missing link in the chain of correlations in the form of a hypothesis. That is, he makes a set of assumptions, or in other words, he invents some imaginary "facts" which provide a causal connection between A and C, *i.e.*, make C appear as a logical consequence of A or vice versa or both A and C appear as a logical consequence of a common cause. This is a powerful method. It enables one to explain everything by making suitable arbitrary assumptions. But this method must be used with caution, since it may introduce into science, in addition to known facts, a large number of questionable assumptions. If some of these assumptions are logically unrelated, the scientist will proceed to invent hypotheses to correlate hypotheses. It is easy to see where such an indiscriminate generous policy in regard to the invention of hypotheses would lead. It is, therefore, of primary importance to keep the number of our hypotheses at a minimum. As a rule, in adopting a hypothesis to link phenomenon A to phenomenon C we shall try to choose a hypothesis which also links A and C to phenomena D, E, F, etc., in other words, to the largest possible number of phenomena, and which is not inconsistent with any of the known facts.

If a hypothesis accomplishes no more than to provide logical continuity between facts, its value is rather limited. It is then merely a useful device in systematizing knowledge which aids our memory. **A fruitful *hypothesis* should suggest the existence of yet undiscovered facts. Such predictions can be tested by experiment. Every new successful prediction of this kind strengthens our belief in the correctness of the set of assumptions underlying the hypothesis which, after a convincing amount of verified predictions, is dignified by the name of *theory*.**

The final step in confirming our belief in a theory is a direct experimental verification of the existence of all the assumed elements of the hypothesis. For instance, chemists assumed the existence of atoms which were supposed to be arranged in regular geometric patterns to form molecules. The general use of and belief in this theory was justified indirectly by the success of its chemical predictions. But it was not until the advent of X-ray structure analysis that the existence and geometrical arrangement of discrete atoms in molecules was ascertained directly (see Chap. 34).

We shall attempt in Chaps. 19 and 20 to develop a theory explaining the phenomena of dissipation of mechanical energy, thermal behavior of matter, mechanical behavior of gases, and chemical combination of matter by means of a single set of assumptions. In establishing this theory we shall follow the procedure outlined above.

Discontinuity of Matter

We have considered in previous chapters compressibility of liquids and of gases. Imagine, for instance, a cube of water subdivided into a large number of cubical volume elements. In explaining the observed shrinkage of the volume under compression, we shall reject the possibility that an increase in pressure makes the tiny cubical volume elements into which we subdivided our cubical volume penetrate each other. This would violate the postulate that two continuous bodies cannot occupy the same space at the same time. A simple alternative would be to assume that water is not continuous but has a spongelike structure, *i.e.*, contains empty spaces in its interior into which matter can be compressed. Since all matter is compressible, this view suggests a discontinuous structure for all matter. According to this view, if we subdivide a cube of matter into small enough volume elements, some of them will be empty or at least partly so. Are we to imagine water to have a coherent spongy structure or to be more like a powder consisting of very small disconnected particles? The ease with which water can flow and mix with other fluids suggests the latter analogy. If we dissolve sugar in water, we obtain a very intimate mixture. No matter how small a volume of liquid we isolate attempting to separate blobs of water from tiny grains of sugar, we shall fail in our attempts. The mixture is very intimate indeed, as if we had a mixture of two powders of imperceptibly small grain size. Thus we are led to ascribe a discontinuous structure to sugar as well as to water. Since similar considerations on mixing and compressibility can also be made in the case of gases, we *shall make a tentative general assumption that matter in all phases has a granular structure.* We shall call these tiny granules **molecules.**

Molecules and Atoms

Are the molecules of a given substance all alike or of varying size? The answer to this was provided by chemical studies. Chemists discovered that chemical elements always occur in definite weight proportions in their compounds. Dalton suggested that this behavior could be explained by assuming each element to consist of extremely small particles of equal size, called **atoms,** which are not further subdivisible. If we now assume that an atom of fluorine weighs 19 times as much as an atom of hydrogen and that hydrofluoric acid is formed by combination of one H atom with one F atom to form one **molecule** of HF, then we can understand why the weight ratio of fluorine to hydrogen in hydrofluoric acid is 19:1 and why intermediate ratios are impossible.

Molecular Motion

Are the molecules at rest or in motion? A simple experiment suggests that they are in motion. If we fill a cylinder half with water and carefully pour a column of alcohol over it, we shall eventually detect alcohol in the volume of water, while the overlying alcohol will be getting gradually diluted with water. The spontaneous penetration of alcohol molecules past the dividing line between water and alcohol is an evidence of their motion, just as the upward penetration of water into alcohol demonstrates that water molecules, too, are in constant motion.

What will happen if we pour water on top of water in our cylinder? Then it is easy to infer from the preceding experiment that water molecules should be moving both ways across an imaginary boundary. In other words, water molecules at any point of our water volume should be in constant motion.

Diffusion experiments similar to the aforementioned mixing process of water and alcohol are more striking when performed with gases, but even solids may exhibit a very slow diffusion. Thus, for instance, diffusion of gold into lead and of lead into lead [using isotopic tracer technique (see Chap. 39)] has been demonstrated. We are thus led to the assumption that molecules are in a state of agitation in solids as well as in gases and liquids.

The rate of diffusion increases with temperature. This suggests the assumption that the molecular agitation increases with temperature. If we succeed in associating temperature increase in a quantitative way with increase in molecular agitation, we may be able to arrive at a mechanical interpretation of the concept of temperature.

Our assumption of correlation between temperature and molecular agitation is supported by numerous phenomena in which an increase of temperature is observed and in which we would expect, for mechanical reasons, the molecular agitation to be increased. For instance, when we compress a gas in a cylinder, its temperature goes up. According to our hypothesis, this temperature increase is to be accompanied by increased molecular agitation. Can it be inferred from the laws of mechanics that the process of compression should result in acceleration of molecules? It is relatively simple to show that the molecular agitation would be increased by compression. It is shown in the example at the end of this chapter that, when a moving elastic ball strikes a wall which recedes from it, the ball is reflected in such a way that the magnitude of its velocity is decreased. Similarly, it can be shown that, when the wall moves toward the ball, the magnitude of the ball's velocity is increased by the process of elastic reflection, whereas a ball moving straight toward a stationary wall suffers a reversal in the direction of its velocity but does not change its speed.

If we now accept the following model of a gas, we shall be able to explain, in addition to the adiabatic heating and cooling of gases, a great variety of other phenomena.

Assumptions of the Kinetic Theory of Gases

We make the following assumptions:

1. A chemically uniform gas consists of identical molecules.

2. The molecules are constantly in motion. This motion is entirely chaotic. The individual molecules may have different velocities at any given moment, and one individual molecule may have different velocities when examined at different times, but at constant temperature, a definite constant average speed may be assigned to the group of molecules.

3. The molecules behave like smooth elastic spheres; *i.e.*, they bounce off from stationary walls and from each other in such a way that in each collision process the sum of the kinetic energies and the vector sum of the momenta of the colliding bodies remain conserved.

4. In a greatly rarefied gas, molecules would bounce from wall to wall, collisions between individual molecules being highly improbable owing to their smallness. But at atmospheric pressure their concentration and, hence, the likelihood of intermolecular collisions are very high. Each molecule undergoes numerous collisions with other molecules as it moves along a zigzag path from wall to wall. Even though the length of a straight path between two collisions varies from molecule to molecule and from impact to impact, we can speak of a definite **mean free path** (*i.e.*, average distance traversed between two collisions) for a given gas at a given pressure and temperature.

5. The molecular radii are assumed to be negligibly small as compared with the mean free path.

6. The molecules do not exert forces upon each other (except during the negligibly short period of encounter).

7. The duration of a collision is negligible as compared with the time interval between two collisions.

Adiabatic Compression and Expansion

In our example of a gas in the process of compression, the molecules, which strike the piston as it moves into the cylindrical gas volume, rebound with a higher velocity and, hence, higher kinetic energy. Their gain in kinetic energy is distributed among other molecules with which they collide, and thus by a process of energy distribution by collision, the total content of kinetic energy of the enclosed molecules is raised.

In the converse process of adiabatic expansion the molecules are allowed to do work on the piston (Fig. 18.1). They may, for instance, accelerate it. The piston then gains kinetic energy at the expense of the kinetic energy of the molecules. Since the piston moves in the processes of expansion in the direction in which the impinging molecules are moving, the latter are reflected with a smaller velocity; hence they lose kinetic energy, and the gas is cooled.

The observable temperature change is seen to be associated with a hypothetical change in average kinetic energy of gas molecules, and we can suspect (as did Lavoisier and Laplace in the nineteenth century!) that the average kinetic energy of the molecules is a measure of temperature.

Conduction of Heat

Our preceding consideration also suggests a mechanical picture of heat conduction. The distribution of kinetic molecular energy by collision over the entire gas volume can be accepted as a picture of the distribution of heat over a gas volume, where the faster molecules of the hotter section strike the slower molecules of the cooler section and lose part of their energy to them until we have the same average kinetic energy in any selected small sample volume of our total gas volume, *i.e.*, a uniform temperature. This picture (suitably modified) can be extended to explain conduction in liquids and solids.

Conversion of Mechanical Energy into Heat

We can now understand why a bullet which strikes the target gets hot and heats the target. Each molecule of the moving bullet has superimposed upon its instantaneous velocity of **disorderly** (or random) molecular motion a component of **orderly** motion in the direction of flight of the bullet. As the bullet strikes the target, this orderly motion ceases. What happened to the energy $\frac{1}{2}mv_b^2$ which each molecule[†] of the bullet possessed in addition to its kinetic energy of disorderly molecular motion? The sum of the kinetic energies of orderly motion of the molecules was distributed over the molecules of the bullet and of the target by intermolecular elastic impacts. *The kinetic energy of orderly motion was converted into kinetic energy of disorderly molecular motion.* Avoiding a terminology based on a hypothesis we can describe the same facts by stating: *The kinetic energy of the bullet was transformed into heat.* We thus see that we can **interpret heat as mechanical energy of the molecules of a body at rest.**

[†] m = mass of the molecule; v_b = velocity of the bullet with respect to the target.

Mechanism of Generation of Gas Pressure

The same picture which enabled us to explain the conversion of mechanical energy into heat also explains why a gas exerts a pressure and why this pressure changes with volume and temperature.

Imagine a horizontal plate released to fall. During 1 sec the force of gravity will have exerted upon it the impulse $Ft = W \times 1$ lb-sec, and as a result of this impulse, the plate has acquired the final momentum: $mv = (W/g)v = W \times 1$. Could we have prevented the plate from falling without suspending it from strings or supporting it by some solid support from below? The answer is yes. We could have bombarded it from below with a battery of peashooters, imparting to it a series of impulses to neutralize the impulse due to gravity. If v is the velocity with which a pea strikes the hovering plate from below, it will be reflected with a downward velocity $-v$. The change of velocity on reflection is $\Delta v = v - (-v) = 2v$, and the change of momentum of an individual pea is $m\,\Delta v = 2mv$. This change of momentum of a pea must be equal to the impulse it imparts to the plate [equation (5.2)]. If we shoot n peas per second, the sum of their impulses over a second will be $n2mv$. The same impulse could have been produced by exerting a constant force F for the duration of 1 sec. This force would exert an impulse $I = F \times 1$ in 1 sec, and we can consider this impulse as being equivalent to the impulses which were delivered in a second by the peas to the plate. Thus we can compute the average force which the peas exert upon the plate by a statement equating the sum of the impulses of the peas in a second to the impulse of the average force: $F \times 1 = n2mv$, or

$$\underline{F = 2nmv} \tag{19.1}$$

If $F = W$ (the weight of the plate), the plate will be prevented from falling. The discontinuous bombardment may make the plate dance a bit up and down, but we can achieve a smoother performance by reducing the size (or velocity) of our peas and shooting more of them per second. By replacing the peas by molecules and the plate by the wall of a container, we obtain the explanation of the origin of gas pressure according to the kinetic molecular theory.

It is easy to see why the pressure is doubled when the gas volume is halved. The molecules will be concentrated in such a way that twice as many of them will be striking the wall each second (Boyle's law).

It is also easy to understand why the pressure should increase with temperature. A rise of temperature, while increasing the average kinetic energy, will also increase the average momentum of the molecules; hence, the average impulse imparted to the wall by reflected molecules becomes larger. In addition, collisions with the wall become more frequent owing to increased molecular velocities.

Similarly, the increase in volume at constant pressure can be explained as follows: A temperature increase of a given volume of gas raises the pressure. To maintain constant pressure the volume has to be enlarged in order to reduce the number of molecules striking the wall each second.

The hypothesis of molecular agitation also explains why a gas has the tendency of filling any given space, why one gas tends to diffuse into another, and why the rate of diffusion increases with temperature.

Absolute Zero

Can we reduce the temperature of our gas indefinitely? Scarcely. We can expect to be able to cool the gas at best until its molecules have stopped moving. Thereafter, according to our theoretical conception of temperature, no further temperature reduction could take place. At this point the gas pressure should drop to zero. Thus, the kinetic theory predicts the existence of an "absolute zero."

Intermolecular Forces

If the molecules of liquids and solids move similarly to gas molecules, why doesn't Boyle's law also apply to liquids and solids? Let us remember that one of the assumptions on which the kinetic molecular theory of gases is based (assumption 6) states that forces of intermolecular attraction or repulsion can be neglected. Whereas for gases this assumption can be upheld, as we shall see presently, it is easy to see that it does not hold at all for liquids or solids. Molecules of solids and liquids are held together by attractive forces of cohesion. For instance, when we suspend a weight from a steel bar, the latter is slightly elongated. The molecules of the solid must have moved somewhat apart. The force of reaction which keeps the weight suspended must be provided by forces of intermolecular attraction.

Conversely, if we try to compress a solid or a liquid, we succeed in doing so, but the compression achieved by even a large force is very minute as compared with the compression of a gas under similar conditions. Figure 19.1*a* shows the volume of a gas as a function of pressure (at constant temperature). Figure 19.1*b* shows the volume of a liquid as a function of pressure (at constant temperature). We notice how much less compressible the liquid is. A very slight diminution in volume requires an enormous increase in pressure. Similarly, a very flat curve would be obtained for the compression of a solid. For instance, if we place a weight on top of a steel rod, the latter is slightly compressed and the force of reaction which balances the weight must be provided by forces of intermolecular repulsion. But how are we to understand this?! The same molecules of the steel rod which exhibit mutual attraction when we apply tension to the rod show mutual repulsion when we try to compress the rod and do not repel or attract each other when neither pressure nor tension is applied to the rod!

This can be explained by the following hypothesis: Let us consider two representative neighbor molecules of the steel rod. They must be normally in a state of equilibrium; otherwise they would move toward each other or away from each other with acceleration. Let us assume that each molecule exerts upon the other both a force of attraction and a force of repulsion. Let us further assume that each of these forces diminishes with the intermolecular distance r according to a law of the form $F = K/r^n$, where K is a constant which may be different for the law of attraction and the law of repulsion. We assume, furthermore, that n has a larger value for the law of repulsion (ranging in general from 9 to 14, depending on the molecular species considered) than for the law of attraction (ranging between 5 and 6). This means that the force of repulsion diminishes with the intermolecular distance much more rapidly than the force of attraction. In Fig. 19.2 the dashed curve shows forces of attraction which diminish with a relatively low power of the distance r. Forces of repulsion are plotted in the direction of the negative Y axis, being opposite in direction to the attractive forces. They diminish much more steeply with the distance as shown by the dotted curve. The resultant force acting on a molecule is obtained by adding the attractive and the repulsive forces (addition of ordinates in our figure). We obtain thus the solid curve for the resultant force. *We see that there is a distance r_0 at which the resultant force is zero. This distance determines the equilibrium position of the molecules.* When we compress the two molecules toward each other ($r < r_0$), the force of

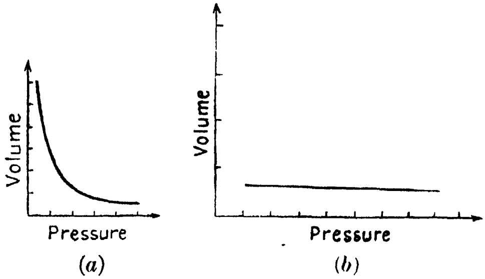


FIG. 19.1*a*. Isothermal relation between pressure and volume for an ideal gas. *b*. Isothermal relation between pressure and volume for a typical liquid.

repulsion increases more rapidly than the force of attraction (since $n > m$ in Fig. 19.2), and the result is a force of repulsion which tends to separate the two molecules until they assume their equilibrium distance. If we try, conversely, to pull the molecules apart ($r > r_0$), the force of repulsion diminishes more rapidly than the force of attraction and we obtain a resultant force of attraction which tends to approximate the molecules toward each other until they reach the equilibrium separation of r_0. It is easy to visualize how a disturbance of the equilibrium between the two molecules will lead to an oscillation of the molecules. In fact, the molecular motion in the case of solids is pictured as an irregular vibration of the molecules about an equilibrium position. The conditions in liquids are as yet not well understood, but we can tentatively visualize the molecular motion in them like the motion of jittery little worms creeping in all directions in a dense heap.

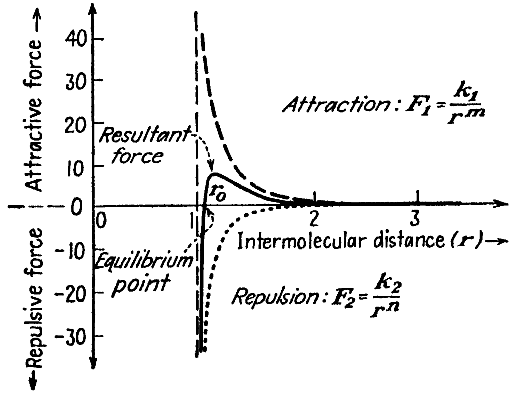


FIG. 19.2. Intermolecular force as a function of intermolecular distance.

After this consideration one may well ask the question: Was it necessary to adopt the kinetic molecular hypothesis in order to explain Boyle's law? Could it not have been explained more simply by assuming repulsive forces between the molecules which would account for the tendency of a gas to expand? As a matter of fact such a hypothesis was considered before the kinetic theory of gases but was abandoned after it proved impossible to deduce from it Boyle's law and to predict other properties of gases.

We mentioned at the outset of this discussion that one of the assumptions of the kinetic gas theory is that gas molecules exert no forces upon each other (except at the moment of collision). How can this assumption be reconciled with the existence of the intermolecular attractive forces which have been demonstrated? How do we reconcile the fact that molecules of gaseous air do not seem to attract each other whereas those of liquid air do? When we liquefy air, its volume is greatly decreased. We assume that the molecules move close enough together to exert upon each other attractive forces which we have previously assumed (in Chap. 11) to explain surface tension. In the gaseous phase the same molecules are too far apart to exert upon each other noticeable forces. It is true that in a gas molecules come close enough together to repel each other in the process of collision, but the duration of collision is so small in comparison with the duration of "free flight" outside the range of intermolecular forces that we disregard the latter and consider the entire internal energy of a gas to be exclusively due to the kinetic energy of its molecules.

Evidence Supporting the Assumption of Intermolecular Attraction

What evidence do we have to support the assumption of mutual attraction between the molecules? Or better, what could we predict about the behavior of matter, assuming close-range intermolecular forces of attraction?

1. If we compress a gas, it should first follow Boyle's law, but at high enough pressure it should become more compressible than predicted by Boyle's law owing to the mutual attraction of the molecules. This is precisely the behavior predicted by van der Waals' equation and verified by experiments. [The term a/V^2 in equation (18.5) is due to intermolecular attraction.]

2. If we compress a gas very strongly at a low temperature, the molecules will be spending an appreciable amount of time in the attractive region of the field of their neighbor molecules. If we now allow the gas to expand freely in an empty container, the gas should cool off, because molecules will be receding from each other moving against forces of mutual attraction and thus losing speed. This is the explanation of the **Joule-Thomson effect** which we described in Chap. 18.

3. *Cooling by evaporation* can be explained similarly: First of all, the fastest of the molecules on the liquid surface will tend to escape more frequently than the slower ones. Depletion of the stock of fast molecules will lower the average kinetic energy of the remaining liquid molecules; in other words, the temperature of the remaining liquid will be lowered. Does this mean that the escaping vapor will be hotter than the liquid? By no means! The escaping molecules of vapor will be slowed down by the attractive forces of the surface molecules as they move away from the surface. By this process the escaping vapor is cooled precisely to the temperature of the liquid which it left. The *heat of evaporation* is the energy required to overcome these attractive forces in the process of the escape of vapor molecules from the liquid surface.

4. Conversely, gas molecules which move toward the liquid surface will be speeded up by attraction toward the surface. The gain in their kinetic energy gives rise to the liberation of *heat of condensation*, which is equal to the heat of evaporation.

Dynamic Equilibrium between Liquid and Saturated Vapor

In the case of a saturated vapor in equilibrium with its liquid there should be a constant stream of molecules leaving the liquid surface and an equal stream of vapor molecules moving toward the surface and condensing. A "compression" of the vapor volume will concentrate the vapor and thus increase the concentration of molecules returning to the surface. When the surplus molecules have condensed, the rate of condensation of molecules will be again equal to the rate of evaporation, which is determined by the prevailing temperature.

Brownian Motion

The kinetic molecular theory would be rendered more convincing if we could supply some directly visible evidence of the actual motion of the molecules. Let us recall our peashooter experiment. If the plate were sufficiently light and the number of impinging peas small enough, the irregular bouncing of the plate would reveal the fact that it is held up by the impacts of peas. Could we not do something similar with a gas? Could we perhaps build a tiny cylinder with a very small weighted piston and rarefy the gas in the cylinder until the irregular motion of the piston will reveal the discrete impacts of the molecules?

A modification of such an experiment was actually performed unintentionally. The botanist Brown observed that tiny spores of plants were performing a never-ceasing jittery dance in the field of his microscope. He thought at first that it was spontaneous motion of living organisms. But later investigations showed that any small particles, solid or fluid, suspended in a liquid or a gas move about in a jerky, irregular fashion. The smaller the particle the greater the mobility. This so-called **Brownian motion** was finally interpreted as being caused by the impacts of the molecules of the fluid against the tiny particles suspended in it. These particles play the role of the minute pistons of our projected imaginary experiment. They are so small (less than 10^{-4} cm in diameter) that the discontinuous mechanism of pressure exertion becomes apparent. The resultant impulse due to impacts of molecules hitting our test particles is not exactly zero over a short period of time. Sometimes a particularly fast molecule may strike the tiny particle and change its momentum; sometimes more impacts will be delivered from the right than from the left side. The resulting Brownian motion is an exact replica of molecular movements. The suspended particles merely behave like giant molecules, which, owing to their large size, move more slowly. But the cause of their motion is the same as the cause of molecular motion, namely, impacts of neighbor molecules.

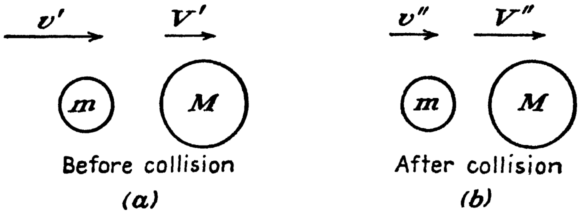
The displacement in Brownian motion depends at a given temperature for a given particle upon the viscosity of the fluid. It has proved possible to estimate changes in viscosity in living cells by observing the variation in the Brownian movement of tiny granules suspended throughout the volume of the cell.

The study of Brownian motion provided, as we shall see in Chap. 20, not only a qualitative but also an excellent quantitative confirmation of the predictions of the kinetic molecular theory.

Example

An oxygen molecule moving west with the velocity of 10^3 m/sec strikes the wall of a piston which moves westward with the speed of 50 m/sec. With what velocity is the molecule reflected from the wall?

Preliminary Consideration. Let us first consider the general case of a collision between two perfectly elastic spheres and derive the theorem that *in central impact*† *the relative velocity between the spheres after the collision is equal and opposite to the relative velocity before the collision.*



Before collision (a)   After collision (b)

In any kind of impact, momentum is conserved. This is expressed by the conservation equation

$$mv' + MV' = mv'' + MV'' \qquad (a)$$

(Vector sum of momenta before impact = vector sum of momenta after impact)

In a perfectly elastic collision the kinetic energy remains conserved:

$$\frac{m}{2}(v')^2 + \frac{M}{2}(V')^2 = \frac{m}{2}(v'')^2 + \frac{M}{2}(V'')^2 \qquad (b)$$

$$\begin{pmatrix}\text{Kinetic energy of the}\\ \text{spheres before impact}\end{pmatrix} = \begin{pmatrix}\text{kinetic energy of the}\\ \text{spheres after impact}\end{pmatrix}$$

Multiplying equation (*b*) by 2 and factoring out m and M, we get

$$m[(v')^2 - (v'')^2] = M[(V'')^2 - (V')^2] \qquad (c)$$

---

† In "central impact" the velocity vector of the impinging sphere is parallel to the line joining the centers of the spheres.

Since $a^2 - b^2 = (a + b)(a - b)$, we can rewrite equation (*c*) as follows:

$$m(v' + v'')(v' - v'') = M(V'' + V')(V'' - V') \tag{d}$$

Equation (*a*) can be written as follows:

$$m(v' - v'') = M(V'' - V') \tag{a'}$$

Dividing equation (*d*) by equation (*a'*), we obtain

$$(v' + v'') = (V'' + V') \tag{e}$$

This can be written as follows:

$$\boxed{(v' - V') = -(v'' - V'')} \tag{f}$$

$$\begin{pmatrix}\text{Relative speed of approach}\\ \text{between the spheres}\end{pmatrix} = \begin{pmatrix}\text{relative speed of recession}\\ \text{between the spheres}\end{pmatrix}$$

Solution. We treat the collision of the molecule of mass m with the piston of a practically infinitely larger mass M like a central collision between two spheres. Equation (*f*) holds independently of the values of the masses. Hence we can write for our case where $v' = 10^3$ m/sec, $V' = 50$ m/sec, $V'' = 50$ m/sec

$$(1{,}000 - 50) = -(v'' - 50)$$

Hence, $$v'' = 50 + 50 - 1{,}000 = -900 \text{ m/sec}$$

We see that, whereas the magnitude of the relative speed *with respect to the moving piston* did not change [this is the condition implied in equation (*f*)], the speed *relative to the room* has diminished from 1,000 to 900 m/sec.

The molecule is seen to have lost kinetic energy. What happened to this energy? Were we justified in treating this case as an elastic impact? What light does this example shed on the mechanism of adiabatic cooling by expansion?

Show that the molecule would have gained kinetic energy if the piston were moving in the opposite direction.

QUESTIONS AND PROBLEMS

1. How does the molecular motion in water at 0°C differ from the molecular motion in ice at 0°C? (Give reasons for your assumption.)

2. A ball moving east with a velocity of 100 ft/sec, strikes a wall which is moving east with a velocity of 20 ft/sec. In what direction and with what horizontal velocity does the ball move after an elastic collision with the wall? (The wall behaves like a body of infinite mass and does not change its velocity after collision.)

3. Solve Prob. 2 assuming the wall to move westward instead of eastward.

4. How fast and which way should the wall move in order to stop the ball by collision? (Measure all velocities with respect to ground.)

5. Ball A, moving east with the velocity of 100 ft/sec, strikes a stationary ball B of equal mass. What are the velocities of A and B after an elastic, central collision?

6. Why does (*a*) adiabatic compression heat a gas? (*b*) adiabatic expansion cool it?

7. How large is the change of momentum of a ball reflected elastically from a wall? (Resolve the momentum vector into two components, one parallel to the wall and the other perpendicular to the wall.)

8. Deduce how a ball striking a plane wall at the angle of θ degrees against the surface will be reflected.

9. What evidence suggests the existence of (*a*) intermolecular attraction? (*b*) intermolecular repulsion?

10. What is the energy supplied to a liquid boiling at constant temperature used for according to the kinetic molecular theory?

11. Is the entire molecular energy present in the form of kinetic energy (*a*) in a gas? (*b*) in a solid?

12. Can intermolecular attraction be due to gravitational forces?

13. Why don't liquids obey Boyle's law?

14. What is the physical significance, according to the kinetic molecular theory, of the term added to the pressure p in van der Waals' equation (18.5)?

15. How does the kinetic molecular theory explain the Joule-Thomson effect?

16. How does the kinetic molecular theory explain heat of condensation?

17. How would the character of the Brownian movement change if we were to decrease the size of the suspended particles?

18. How would it change if we were to heat the fluid?

19. What happens to the intermolecular forces in a solid rod (*a*) when it is compressed? (*b*) when it is put under tension?

20. Do molecules of a solid attract each other or repel each other? On what evidence do you base your answer?

CHAPTER 20

KINETIC MOLECULAR THEORY: 2. QUANTITATIVE PART

a. How do we know that the laws of mechanics apply to the unseen molecules when we apply these laws to the kinetic molecular model of a gas?

b. What is the mean free path of a molecule?

c. How do you determine experimentally the probability of a marble's landing in one of the compartments of Galton's board (Fig. 20.1)?

d. What does the numerical probability determined in *c* tell you?

e. Can you predict whether or not a marble released in the setup of Fig. 20.1 will land in a definite compartment?

f. What could you predict quite accurately in regard to marbles released in the setup of Fig. 20.1?

g. What are accidental events?

h. Can you explain the significance of the ordinates of the Gaussian error curve (Fig. 20.2)?

i. Why don't all molecules move equally fast?

j. What is the root mean square velocity of gas molecules? Why do we consider it instead of a simple average velocity?

k. How does a temperature increase affect the shape of the Maxwellian velocity distribution?

l. What simplifying assumptions did we make in deriving Boyle's law, and how did we justify them?

m. From what observable data can we compute the rms velocity of the molecules?

n. From what premises did we deduce the general gas law?

o. What relationship exists between the rms velocity of the molecules and the absolute temperature of a gas?

p. From what premises did we deduce Avogadro's law?

q. What does Avogadro's law state?

r. Why does a dissolved substance exert an osmotic pressure equal to the gas pressure it would have exerted as a gas at the same temperature and concentration?

In Chap. 19 we presented a qualitative picture of the constitution of matter based on the assumptions of the kinetic molecular hypothesis in an attempt to explain the behavior of liquids and particularly of gases in a nonquantitative manner. The purpose of the hypothesis was to associate an imaginary mechanical process with the phenomena of evaporation, condensation, adiabatic and isothermal volume changes, conversion of mechanical energy into heat, Joule-Thomson effect, etc., in order to correlate new knowledge with the well-established laws of mechanics.

In this chapter we shall apply the laws of mechanics to the hypothetical molecules in order to see whether we can deduce quantitatively the gas laws from the basic assumptions of the kinetic molecular theory.

Thereafter we shall attempt to compute the molecular dimensions from experimental data on the basis of the new versions of the gas laws expressed in terms of the concepts of the kinetic molecular theory. And finally we shall check some of the theoretically derived values by direct experimental verification.

Order of Magnitude of Molecular Data

The following is a survey of molecular magnitudes as they can be computed from experimental data on the basis of kinetic molecular theory. This list is intended to give us an idea of the *order of magnitude* of the molecular quantities with which we are dealing.

Radius of an O_2 molecule	1.3×10^{-8} cm.
Mass of an O_2 molecule	52×10^{-24} gm.
Average speed of O_2 molecules at standard p and T	425 m/sec.
Number of molecules in 1 cm^3 of air at standard p and T	27×10^{18} molecules.
Mean free path of O_2 molecules at standard p and T	10^{-5} cm.
Number of collisions per second suffered by an O_2 molecule at standard conditions of p and T	4165×10^{6} per second.

The radius and mass seem hopelessly small to be measured, and the number of molecules per cubic centimeter would take about 500,000 years to count if we could count individual molecules. Despite these handicaps, ingenious methods enabled us to determine these quantities with relative ease.

THEORETICAL DERIVATION OF GAS LAWS

Statistical Distributions

Let us begin with the following preliminary consideration: Figure 20.1 shows an arrangement for the production of accidental events (Galton's board).

It consists of a vertical wooden board studded with projecting nails arranged in rows as shown in the figure. Small spheres which are just able to slip through the opening of the funnel hit the uppermost center nail in falling and are deviated to the right or to the left and continue on an *unpredictable* zigzag path, bouncing from nail to nail until they land in one of the storage compartments at the bottom. The path is unpredictable not because the laws of mechanics do not hold in this case but because very slight variations in the initial linear and rotational velocities of the ball, which we cannot practically ascertain, determine its course. Cases determined by factors which we cannot ascertain or control in practice are called **accidental events.** The word accidental does not, however, imply that the processes don't proceed according to the known laws of physics; it merely implies our ignorance of the details of an individual process.

To illustrate how consideration of accidental events may be involved in an analysis of a physical situation let us consider the following example:

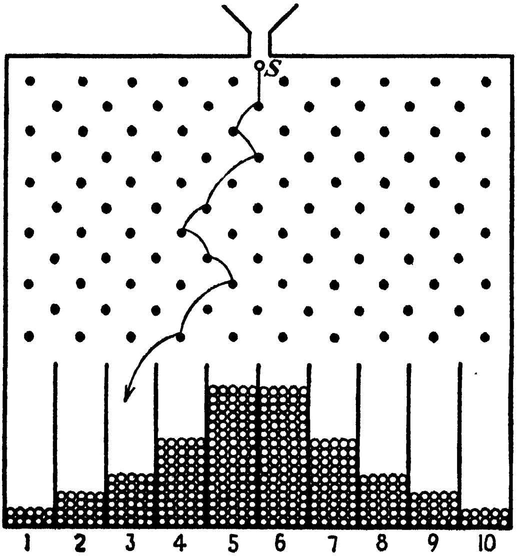


FIG. 20.1. Galton's board.

We could, for instance, devise a game of chance assigning different numbers to the compartments of Fig. 20.1 and rewarding or penalizing a player according to where the sphere which he releases lands. To make it a profitable business we shall impose a high penalty in case the sphere lands where it is most likely to land and offer a generous reward in case the sphere lands in a place where it is hardly ever deposited. In our decision as to how to distribute our penalties and awards we shall be guided by the results of many experiments. If we release a great many spheres, they will accumulate in our numbered compartments, forming a heap of a regular, symmetrical outline. After the release of only a few spheres, no apparent regularity will be noticeable but the regularity and smoothness of the contour of the heap will increase with the growing number of spheres deposited.

The above experiment suggests a simple method of defining the probability (likelihood) that a sphere may land in, say, compartment 8. Our intuition tells us that this likelihood is measured by the percentage of the released spheres which land in compartment 8. Thus the probability of a released sphere being trapped in 8 is $\frac{\text{number of spheres found in 8}}{\text{total number of spheres released}}$.

The result of our experiment indicates that a sphere is most likely to get into boxes 5 or 6 and least likely into 1 and 10. The contour of the heap can be called a **probability distribution** curve; it tells us how the probability of capturing a falling sphere is distributed over the different compartments.

This curve is identical with the Gaussian error curve which depicts the distribution of errors in measurements. When a skillful physicist measures a certain physical quantity, he does not always "hit" the right value precisely. Most of the time he is close, but occasionally, owing to accidental causes beyond his control, his readings may be appreciably off. If we plot the number of times he repeated a certain reading vertically above the corresponding mark on the scale, we shall get Fig. 20.2. If the correct value of the quantity he measures is 3, it will be the most frequent reading he obtains; it will be his "most probable" reading.

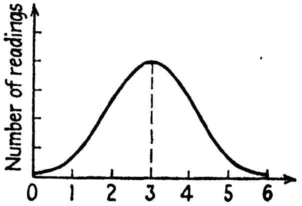


FIG. 20.2. The Gaussian error curve.

The Maxwellian Velocity Distribution

Maxwell showed that the velocity cannot be the same for all the molecules of a gas. All values of velocity from zero to exceedingly high values are possible but not equally probable. The probability function for the distribution of the molecular velocities which he derived was shown by Boltzmann to be related and similar to the Gaussian error function. There is a certain most probable velocity v^* with most of the molecules moving at speeds close to v^*, but velocities near zero or exceeding the most probable velocity many times are possible.

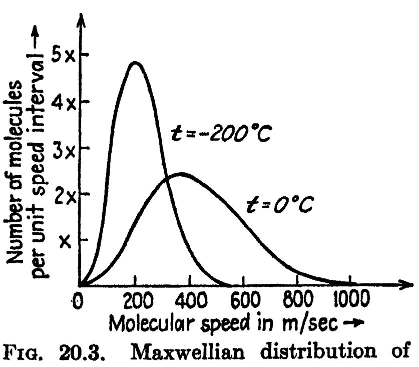


FIG. 20.3. Maxwellian distribution of molecular velocities.

Figure 20.3 shows the Maxwellian velocity distribution for oxygen molecules at two different temperatures. We see how an increase in temperature brings about an increase of the value of the most probable velocity but diminishes the percentage of the particles which move with approximately that speed. At the same time the total number of particles whose velocity exceeds a critical value (necessary for the initiation of a chemical reaction) increases as compared with the number of slower particles. This increase in the abundance of fast, "energetic" molecules explains the very rapid increase in the rate of chemical reactions with rising temperature as well as the decrease of the metabolic rate of cold-blooded animals at low temperatures.

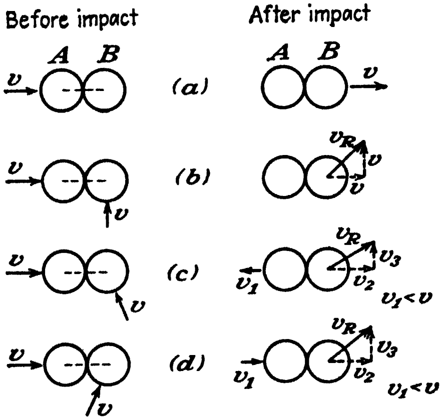


FIG. 20.4. Various cases of elastic impact.

At first sight the Maxwellian velocity distribution does not seem plausible. One might ask the question: If all molecules are identical, what should make molecule *A* move faster than molecule *B* and slower than molecule *C*? A simple consideration shows that a state where all molecules were moving permanently with equal speed would be impossible under the basic assumptions of the kinetic molecular theory. If we had succeeded somehow in imparting to all gas molecules the same speed at one instant, this state would not persist and a velocity redistribution would begin immediately according to the following mechanism (Fig. 20.4):

When a smooth, elastic sphere *A* strikes centrally another elastic sphere *B* of equal mass initially at rest, then *A* stops while *B* proceeds with the original velocity of *A* (Fig. 20.4*a*). If sphere *B* is moving at right angles to *A* (Fig. 20.4*b*) and if at the moment of the impact the line joining the centers of *A* and *B* coincides with the direction of the velocity of *A*, *A* stops and *B* acquires a resultant velocity $v_R > v$. If *B*, now moving faster than originally, should be struck again at right angles as before by a particle *C* moving with the velocity *v*, it will again gain speed and continue to move with the velocity $v'_B = \sqrt{v_B^2 + v^2}$. Thus by successive impacts at right angles to the velocity of B (the line joining the centers of the spheres being perpendicular to v_B), its velocity can be increased indefinitely. Such a sequence of successive accelerating impacts is unlikely but possible. As a rule, there will be interspersed decelerative impacts. Figure 20.4*b* shows how an impact may reduce the velocity of a sphere A to zero, and Figs. 20.4*c* and *d* show other simple cases in which the magnitude of the velocity of one of the colliding bodies is reduced. Thus the incessant collisions between the molecules effect a spontaneous adjustment of a Maxwellian velocity distribution in a gas.

Assumptions Regarding the Velocity Distribution in Deriving Boyle's Law

We are now ready to consider the mechanism of the generation of gas pressure. We discussed this problem qualitatively in Chap. 19 and reached the conclusion that the effect of a continuous pressure can be produced by a rapid succession of molecular impulses. Let us consider a very simple case to begin with. Our gas volume is to contain only one molecule, which bounces back and forth between two walls W_1 and W_2 of area A (Fig. 20.5). The rapid rate of reflection may create the impression of a continuous pressure upon the walls.

We found in Chap. 19 that the impulse imparted by each impact is equal to $2mv$. The total impulse imparted during 1 sec. by n collisions is $F \times 1 = n2mv$ [equation (19.1)] (where F is the average force exerted upon the wall). How many times does the molecule strike a wall in our example in a second? The molecule moves v centimeters in each second. The distance traversed between two successive impacts upon the same wall is $2d$; hence, there should be $n = v/2d$ impacts per second against each wall. Hence, $F = n2mv = (v/2d)2mv = (2/d)\frac{1}{2}mv^2$, and since our walls are to be of area A, we can write

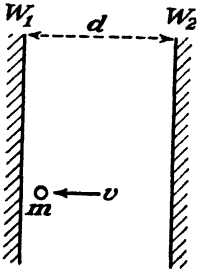


FIG. 20.5.

$$\underline{p = \frac{F}{A} = \frac{2}{Ad}\frac{1}{2}mv^2} \quad . \qquad (20.1)$$

Interestingly enough, the *pressure is proportional to the kinetic energy of the molecule.*

Now assume that we have two molecules of the same mass but different velocity moving in our example. Then

$$p = \frac{2}{Ad}\left(\frac{1}{2}mv_1^2 + \frac{1}{2}mv_2^2\right) = \frac{m}{Ad}(v_1^2 + v_2^2) \qquad (20.2)$$

and for n molecules:

$$p = \frac{2}{Ad}\left(\frac{1}{2}mv_1^2 + \frac{1}{2}mv_2^2 + \frac{1}{2}mv_3^2 + \cdots + \frac{1}{2}mv_n^2\right) = \frac{2}{Ad}(\text{KE})$$

$$= \frac{m}{Ad}(v_1^2 + v_2^2 + v_3^2 + \cdots + v_n^2) \qquad (20.3)$$

From equation (20.2) we see that it makes no difference how the kinetic energy or the velocity is distributed among the two molecules as far as pressure is concerned. The pressure is the same as long as the sum of the kinetic energies (KE) remains the same. From equation (20.3) we see that the pressure produced by a gas under described circumstances will remain unchanged if we vary at will the *velocity distribution among the molecules as long as we maintain the sum of their kinetic energies constant.* Since the average kinetic energy $\overline{\text{KE}}$ of n particles is obtained by adding their kinetic energies and dividing by n, we can write

$$\overline{\text{KE}} = \overline{\left(\frac{1}{2}mv^2\right)} = \frac{\frac{1}{2}mv_1^2 + \frac{1}{2}mv_2^2 + \cdots + \frac{1}{2}mv_n^2}{n} \qquad (20.4)$$

Equations (20.3) and (20.4) enable us to state:

The pressure remains the same regardless of the velocity distribution as long as the average kinetic energy of the molecules remains constant. *Thus, we would obtain the same pressure if we were to assume that all molecules move with the same velocity* (which they don't!), the so-called **rms velocity** $\bar{\bar{v}}$. The value of $\bar{\bar{v}}$ follows from equation (20.4) by setting the kinetic energy of a molecule moving with this fictitious rms velocity, equal to the average of the kinetic energies of all the molecules:†

$$\frac{1}{2}m\bar{\bar{v}}^2 = \frac{1}{2}m\,\frac{v_1^2 + v_2^2 + v_3^2 + \cdots + v_n^2}{n}$$

Hence, the rms velocity is

$$\boxed{\bar{\bar{v}} = \sqrt{\frac{v_1^2 + v_2^2 + v_3^2 + \cdots + v_n^2}{n}}} \qquad (20.5)$$

The rms (root mean square) speed is obviously not the same as the

---

† This is equivalent to the statement that the total kinetic energy of n molecules is equal to n times the energy of a molecule moving with the rms velocity:

$$n\left(\frac{1}{2}m\bar{\bar{v}}^2\right) = \frac{m}{2}v_1^2 + \frac{m}{2}v_2^2 + \frac{m}{2}v_3^2 + \cdots + \frac{m}{2}v_n^2$$

average speed $\bar{v} = \frac{v_1 + v_2 + v_3 + \cdots + v_n}{n}$, nor is it to be confused with the *most probable speed* (velocity at peak in Fig. 20.3).

We shall now make use of the above consideration and compute the gas pressure exerted on the walls of a container under the assumption of equal velocity (equal to the rms velocity) for all molecules.

Derivation of Boyle's Law

In addition, we shall assume that molecules move from wall to wall without collisions. This is obviously contrary to the picture we have presented in Chap. 19, but it is a justifiable simplification. As we know, two identical elastic spheres colliding "head-on" merely exchange their velocities (Prob. 5, Chap. 19). Thus, if molecule A moving east at 1 mi/sec collides with molecule B moving west at $\frac{1}{2}$ mi/sec, molecule B continues *east* at 1 mi/sec, whereas molecule A moves as B did originally, *i.e.*, west at $\frac{1}{2}$ mi/sec. But since molecules A and B are indistinguishable, it is immaterial whether A reached the wall "personally" or molecule B "did it for A."†

Let us now consider a cube of volume L^3 filled with a gas containing N molecules. The molecules move in complete disorder but with the same velocity ($\bar{v}$). Since the velocity vectors have all conceivable orientations in space, the velocity of each molecule will have in general an east-west, a north-south, and an up-down component. We can assume that a portion of the total molecular kinetic energy is due to the east-west motion of the molecules, another portion to the north-south motion, and a third portion due to the vertical velocity component. Since there is no preferential direction in space, it is reasonable to assume that the total kinetic energy of the molecules (KE) is equally distributed over these three directions, so that the kinetic energy due to motion in any one of these directions is KE/3.

Let us now consider the impacts of the molecules upon the north wall. Only the north-south component of their velocity will be effective in producing a pressure upon this wall by change of momentum: According to equation (20.3), $p = (2/Ad)(\text{KE})_{NS}$ where $(\text{KE})_{NS}$ is the molecular kinetic energy due to the motion in the north-south direction. This energy, as we just stated, is equal to one-third of the total molecular kinetic energy of our gas content

$$(\text{KE})_{NS} = \tfrac{1}{3}(\tfrac{1}{2}mv_1^2 + \tfrac{1}{2}mv_2^2 + \cdots)$$

---

† In a case where collision is not "head-on" (for instance, Fig. 20.4c), the velocity v_B can be resolved into two components v_x and v_y. Then, it can be shown that an exchange of velocities takes place along the x axis, *i.e.*, along the direction of the line joining the centers of the spheres.

And since we assumed all molecules to move with the same velocity $\bar{v}$, we can write $(\mathrm{KE})_{NS} = \frac{1}{3}N\frac{1}{2}m\bar{v}^2$. From this equation and from equation (20.3) follows

$$p = \frac{2}{Ad}\frac{1}{3}N\frac{1}{2}m\bar{v}^2 \tag{20.6}$$

For our cube $A = L^2$ and $d = L$; hence,

$$p = \frac{2}{L^3}\frac{1}{3}N\frac{1}{2}m\bar{v}^2 = \frac{Nm}{3L^3}\bar{v}^2 \tag{20.6a}$$

Now $Nm = M$ is the mass of the gas content of the cube, $L^3 = V$ the gas volume, and $M/V = D = M/L^3$ the gas density. Hence,

$$\boxed{p = \frac{Nm}{3V}\bar{v}^2 = \frac{D}{3}\bar{v}^2} \tag{20.7}$$

This equation is Boyle's law in the language of the kinetic molecular theory. It can be written

$$\boxed{pV = \frac{N}{3}m\bar{v}^2} \tag{20.7a}$$

For a given mass of a gas at a fixed temperature, N and $\bar{v}$ are constant.

Since p and D in equation (20.7) are measurable quantities, **we can compute from experimental data the value of the hypothetical rms velocity of the molecules.**

For example, for hydrogen at standard conditions,

$$p = 1{,}013{,}000 \text{ dynes/cm}^2$$
$$D = 0.000090 \text{ gm/cm}^2$$

Hence, $$\bar{v} = \sqrt{\frac{3 \times 1{,}013{,}000}{0.000090}} = 184{,}000 \text{ cm/sec}$$

It still remains to be seen whether this molecular velocity is merely a product of our imagination or its existence and value can be verified experimentally.

Derivation of the General Gas Law

Equation (20.7a) can be written as follows:

$$pV = \frac{N}{3}m\bar{v}^2 = \frac{2}{3}N\left(\frac{1}{2}m\bar{v}^2\right) \tag{20.7a*}$$

We see that the product of pressure and volume of a gas is proportional to its internal kinetic energy $N\frac{1}{2}m\bar{v}^2$.

The internal energy of an ideal gas can be also expressed in caloric terms: If we assume the gas molecules to be at rest at the temperature of absolute zero, we shall ascribe zero internal energy to our gas sample in that state. If we further assume the specific heat at constant volume of our *ideal* gas to be independent of temperature, we shall express the internal energy of a gas at temperature T as the amount of heat spent in heating it from absolute zero to T°K: internal energy $= Mc_vT = (Nm)c_vT$ (where M is the mass of the gas considered). If we substitute this value of internal energy for the mechanical expression for internal energy $N\frac{1}{2}m\bar{v}^2$ in equation (20.7a*), we get

$$\frac{N}{3}m\bar{v}^2 = \frac{2}{3}(Nmc_vT) \tag{20.8}$$

or

$$\boxed{T = \frac{1}{c_vm}\frac{1}{2}m\bar{v}^2 \propto \frac{1}{2}m\bar{v}^2} \tag{20.8a}$$

In words: **The absolute temperature of a gas is proportional to the average kinetic energy of the molecules.**

If we now substitute pV from equation (20.7a*) for $(N/3)m\bar{v}^2$ in equation (20.8), we obtain

$$pV = \tfrac{2}{3}Nmc_vT = \tfrac{2}{3}Mc_vT \tag{20.9}$$

or

$$\underline{\frac{pV}{T} = \frac{2}{3}Mc_v} \tag{20.9a}$$

In the special case where we consider 1 gm mole of a substance, M assumes the significance of the molecular "weight" and the product $Mc_v = C_v$ becomes the specific heat per gram mole.

We can thus write

$$\boxed{\frac{pV}{T} = \frac{2}{3}C_v} \tag{20.10}$$

It has been stated in Chap. 18 that C_v, the specific heat per mole, is the same for all monatomic ideal gases. We thus recognize in equation (20.10) the universal gas law.

By comparison with (18.11) [where we choose $n = 1$, since we are considering 1 mole in equation (20.10)] we see that

$$\boxed{R = \tfrac{2}{3}C_v} \tag{20.11}$$

This simple relation between the specific heat per mole and the universal gas constant is verified by experimental data (see end of Chap. 18).

Derivation of Avogadro's Law

We can now proceed to deduce the law of Avogadro from the equation (20.7*a*) of the kinetic molecular theory by means of the following argument about the distribution of kinetic energy among molecules of different gases at equal temperatures. Let us assume that we have two identical volumes V_1 and V_2 of two different gases subjected to the same pressure $p_1 = p_2$ at the same temperature $T_1 = T_2$.

We can write for them

$$p_1V_1 = \frac{N_1}{3} m_1\bar{v}_1^2 \tag{20.7a'}$$

and

$$p_2V_2 = \frac{N_2}{3} m_2\bar{v}_2^2 \tag{20.7a''}$$

Since $p_1V_1 = p_2V_2$, we get

$$\frac{N_1}{3} m_1\bar{v}_1^2 = \frac{N_2}{3} m_2\bar{v}_2^2 \tag{20.12}$$

or

$$N_2 = N_1 \left(\frac{m_1\bar{v}_1^2}{m_2\bar{v}_2^2}\right) \tag{20.12a}$$

Now it can be shown by considerations based on the laws of mechanics and the laws of probability (statistical mechanics) that the molecules of two or more different gases when mixed will continue to exchange energy by impact until the average kinetic energy of molecules of gas *A* has become equal to the average kinetic energy of the molecules of gas *B*, of the gas *C*, or of any other gas. On the other hand, according to our definition of equality of temperature, two or more gases when mixed will eventually assume the same common temperature. Thus the two gases approach a state of equilibrium in which their temperatures and their average molecular kinetic energies are equal.

The average kinetic energy of the molecules of two (or more) gases is the same if they are at the same temperature.

If we apply this conclusion to equation (20.12*a*), we see that the ratio $m_1\bar{v}_1^2/m_2\bar{v}_2^2 = 1$, since $T_1 = T_2$ and, hence,

$$\boxed{N_1 = N_2}$$

This means that **equal volumes of gas at equal pressure and temperature contain the same number of molecules**

(Law of Avogadro)

EXPERIMENTAL VERIFICATION OF THE KINETIC MOLECULAR THEORY

Having strengthened our confidence in the kinetic molecular hypothesis by deducing from its basic assumptions familiar gas laws, we shall now review how values of various molecular magnitudes have been predicted theoretically and checked experimentally.

Experimental Determination of Molecular Velocities

We have learned how to compute molecular velocities from equation (20.7*a*). Could these data be subjected to an experimental test? An ingenious method has been devised by Stern: Figure 20.6 shows a sector-shaped, *highly evacuated*, closed chamber, which can be rotated about an axis that passes through A at right angles to the plane of the paper. At A there is a furnace from which molecules or atoms of the test substance are evaporated at a known temperature. Those molecules which happen to move in the direction toward point P_1 pass through a small hole in the diaphragm D and hit a screen S at a point P_1, where they are deposited and form a spot. If we now rotate the whole chamber clockwise about the axis, the spot shifts from P_1 to P_2. The molecules leaving the diaphragm D will hit the screen S at the point P_2 instead of P_1,† owing to the rotation of the apparatus, which takes place while the particle is moving from the diaphragm D to the screen S. During this time a line joining A to P_2 will be rotated into position AP_1, sweeping through a small angle P_1AP_2. Knowing the number of revolutions per second we can easily compute the time t required to rotate the apparatus through this angle. This time is identical with the time which a molecule takes to move from D to the screen S. Hence, the velocity of the molecule $v = L/t$ can be computed from the measured displacement of the spot from P_1 to P_2, from the known lengths L and R, and from the frequency of revolution. Thus, for instance, by evaporating silver at a temperature of 1200°C, Stern obtained a deflection of the spot of $P_1P_2 = 1.2$ mm at a frequency of revolution of 2,700 rpm. From this he determined a most probable velocity $v^* = 675$ m/sec.‡ The value predicted by the theory for this case is $v^* = 672$ m/sec.

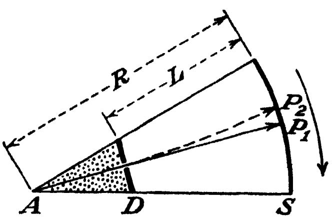


FIG. 20.6. Apparatus for the determination of molecular velocities.

Molecular Diameters and Number of Molecules in a Body

How large are molecules?

To avoid mathematical complications we shall answer this question

---

† The dotted line in Fig. 20.6 represents the path of the molecules as it appears to an observer who is stationed inside the rotating apparatus.

‡ The relationship between the most probable velocity v^* and the rms velocity $\bar{\bar{v}}$ is according to the kinetic molecular theory:

$$\bar{\bar{v}} = v^* \sqrt{\tfrac{3}{2}}$$

by making some rough approximations, which nevertheless will give us a correct idea of the order of magnitude of the molecular diameters.

Let us make the simplifying assumption that molecules are cubical rather than spherical and that, when a gas is solidified, these cubes arrange themselves in an orderly fashion (as shown in Fig. 20.7). Assume that we have a solid cube of side S. If n is the number of molecules contained in the cube, we can express its volume V in terms of the volume of a "molecular cube" as follows:

$$V = nd^3 \qquad (20.13)$$

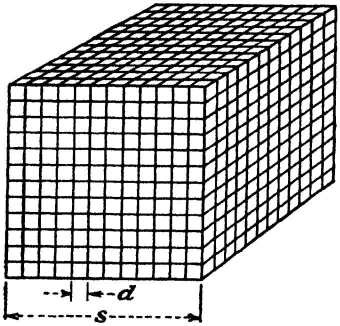


Fig. 20.7.

There are two unknowns in this equation, which are both of interest to us: n, the number of molecules in the given volume, and d, a linear dimension of a molecule. A second equation in n and d is needed to obtain both quantities. Such an equation can be easily set up if we imagine the molecules of our solid cube rearranged to form a *monomolecular layer* (Fig. 20.8). The area covered by this layer of n molecules is

$$A = nd^2 \qquad (20.14)$$

Thus we could determine the size and number of molecules contained in a given mass if we knew only (1) what volume this mass occupies in the solid state and (2) how big an area will be covered by a monomolecular layer of the given mass.

Dividing equation (20.13) by equation (20.14) we get

$$\frac{V}{A} = \frac{nd^3}{nd^2} = \underline{d} \qquad (20.15a)$$

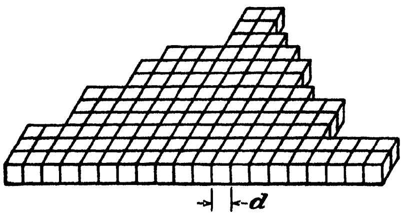


Fig. 20.8.

Substituting equation (20.15a) in (20.14) we obtain

$$\underline{n} = \frac{A}{d^2} = \frac{A}{V^2/A^2} = \frac{A^3}{V^2} \qquad (20.15b)$$

V is quite simple to obtain, but the determination of A presents difficulties. It is possible to produce monomolecular layers of oil on water, for instance, and to measure the area of the layer. But this method is limited to the determination of the size of certain large organic molecules. A method of getting A suitable for the study of gas molecules can be based on the use of molecular beams, as will be shown below.

Suppose we want to determine the average size and number of molecules in a given mass of air with the aid of equations (20.15*a*) and (20.15*b*). The approximate volume V occupied by the molecules of our air quantity, when packed tightly, can be determined from experiments on liquefaction or, better, on solidification of air. But how can we spread air into a monomolecular layer in order to obtain A? This can't be done, but the total cross-sectional area of n molecules can be estimated experimentally as follows:

Imagine silver atoms emanating as a very thin atomic beam from a source S and moving downward like vertically falling raindrops through an evacuated space between the plates D and G, impinging upon a horizontal glass plate G upon which they are condensed (Fig. 20.9). The intensity of the silver beam is measured by determining the amount of silver deposited per unit time on the glass plate. If we now introduce a small amount of air into our vacuum chamber in such a quantity that the sum of the cross-sectional areas of the air molecules is, say, equal to half of the cross-sectional area a of the beam of silver atoms, air molecules will "cast a shadow" on the glass plate; *i.e.*, they will eliminate from the very thin beam those silver atoms which strike them by scattering them out of the beam. Since we admitted only few air molecules into the chamber, it will be very improbable to find two air molecules located behind each other along a vertical line. Thus, the "air shadow" will be cast over the same area as if we had spread the air molecules over the glass plate in form of a monomolecular film in order to obstruct the silver beam. Thus by measuring how much air we have to admit in order to halve the intensity of the silver beam† and by measuring the cross section a of the beam as revealed by the area of the silver deposit, we have all the required data to compute the number and the average size of the air molecules. Let us consider a numerical example:‡

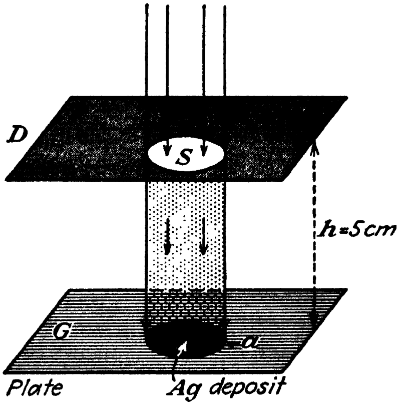


FIG. 20.9.

---

† We make the rough assumption that the intensity will be halved when the total cross-sectional area of the obstructing molecules is equal to one-half the cross section of the beam.

‡ This method of determining d and n and the numerical data are based on a presentation by Professor Born in his "Restless Universe," Harper.

Silver atoms are moving downward after passing through the diaphragm D. We denote the cross-sectional area of the beam by a, the height of the beam by h, the volume of the silver beam by ah.

We find that in order to halve the intensity of the silver beam we have to admit air into the originally evacuated container till the pressure has reached the value $p = p_0/200{,}000$ (where p_0 is the atmospheric pressure). We know from special experiments that a volume V_0 of air at *standard conditions* will be reduced to 1/2,000 of its original value when the air is frozen solid. Thus the air contained in the volume of the silver beam would be reduced to $V^* = ah/2{,}000$ in the solid state if it were originally at 0°C and atmospheric pressure. But we know that the air in our experimental tube is at a very low pressure, namely, $p = 1/200{,}000$ atmosphere. Hence, the solid granule obtained by condensation of the air filling the volume of the beam will be proportionately smaller:

$$V = \frac{ah}{2{,}000}\,\frac{1}{200{,}000} = \frac{ah}{4 \times 10^8}$$

Since the intensity of the beam has been reduced by one-half, the area obstructed by air molecules is $A = a/2$.

Hence, according to equation (20.15),

$$\boxed{d = \frac{V}{A} = \frac{ah}{4 \times 10^8(a/2)} = \frac{\cancel{a} \times 5 \times 2}{\cancel{a} \times 4 \times 10^8} = \frac{1}{4} \times 10^{-7}\ \text{cm}}$$

$$\boxed{d = 2.5 \times 10^{-8}\ \text{cm}}$$

d represents a rough approximation to the average diameter of an air molecule.

Substituting this value of d in equation (20.13) we get

$$V = V^* = nd^3 \quad \therefore \quad V = n(2.5 \times 10^{-8})^3$$

where V^* is the volume of the gas in the solidified state. If we consider 1 cm³ of air at standard conditions, we know that in the solid state it will shrink to a volume of $V = 1/2{,}000$ cm³. Substituting this value in the last equation will give us the number of molecules in 1 cm³ of air:

$$\frac{1}{2{,}000} = n(2.5 \times 10^{-8})^3$$

$$n = \frac{1}{2 \times 10^3 \times (2.5)^3 \times 10^{-24}} = 32 \times 10^{18}\ \text{molecules/cm}^3$$

This is only a rough approximation, based on approximate round figures. Determinations of n based on more precise (and less elementary) methods yield:

$$\boxed{n = 26.9 \times 10^{18} \text{ molecules/cm}^3}$$

(This is the so-called **Loschmidt's number.**) From this value of n we can compute how many molecules there are in 22.4 liters of air at standard conditions, *i.e.*, in 1 gm mole of air and, hence, in 1 gm mole of any substance. We thus obtain the so-called **Avogadro's number.**

$$\boxed{N = 6.023 \cdot 10^{23} \text{ molecules/gm mole}} \qquad (20.16)$$

The Mass of a Hydrogen Atom

The knowledge of Avogadro's number enables us to determine the mass of molecules and atoms. For instance, 1 gm mole of hydrogen weighs 2.0154 gm. Hence, the mass of the diatomic hydrogen molecule should be

$$m' = \frac{2.0154}{6.023 \cdot 10^{23}} = 3.324 \times 10^{-24} \text{ gm}$$

and the **mass of the hydrogen atom**

$$\boxed{m = 1.662 \times 10^{-24} \text{ gm}} \qquad (20.17)$$

Similarly, the other atomic and molecular masses can be determined.

Extension of Ideas of the Kinetic Molecular Theory to Solutions and Suspensions

The ideas of kinetic molecular theory proved also fruitful in the study of solutions and suspensions.

We mentioned in Chap. 18 van't Hoff's law according to which a dissolved substance exerts an osmotic pressure equal to the pressure it would have exerted in the gaseous state at the same temperature and concentration. Why should it do so? The pressure exerted by the molecules of the dissolved substance should depend, at a given concentration, upon the average kinetic energy of its molecules, which is determined by the temperature. If we mix molecules of the variety A with other molecules *at the same temperature,* their average kinetic energies will remain unchanged although there will be a constant interchange of kinetic energy between the molecules A with surrounding molecules through the process of collision. In a solution such a state of "energy equilibrium" exists between the molecules of the solvent and of the dissolved substance. Thus the molecules of, say, sugar cannot lose energy by friction against the water in which they move about. Their average kinetic energy is maintained constant by continuous energy interchange through collisions with water molecules. The presence of the solvent does not interfere with the mechanism of pressure generation by impact, so that the sugar molecules in aqueous solution exert the same pressure as they would have exerted under otherwise equal conditions in a vacuum.

Should then a suspension of sand or pebbles also exert an osmotic pressure similarly to the sugar molecules in solution? Should it not also possess an average kinetic energy per pebble equal to the average kinetic energy of a sugar molecule in solution at the same temperature? We have every reason to assume that they do. Of course it would be hopeless to attempt to measure the osmotic pressure of a suspension of sand, because there would be normally negligibly few sand grains per cubic centimeter as compared with the number of sugar molecules which could be dissolved in the same volume. Should we not at least be able to detect the quasi-molecular motion of the sand grains and pebbles due to impacts of the water molecules? Let us see if this would be possible: Call M the mass of a pebble and $\frac{1}{2}M\overline{\overline{V^2}}$ the average kinetic energy of the pebble due to its "heat motion." The corresponding symbols for the mass and average kinetic energy of, say, oxygen molecules will be m and $\frac{1}{2}m\bar{v}^2$. Then at equal temperature,

$$\tfrac{1}{2}M\overline{\overline{V^2}} = \tfrac{1}{2}m\bar{v}^2$$

and

$$\overline{\overline{V^2}} = \frac{m}{M}\bar{v}^2 \qquad \text{or} \qquad \overline{\overline{V}} = \sqrt{\frac{m}{M}}\,\bar{v}$$

Let us make a rough estimate, using of the data given for oxygen in the first paragraph of this chapter: $m = 52 \times 10^{-24}$ gm and $\bar{v} = 425 \times 10^2$ cm/sec. (It is close enough for our purpose to use the value of the average velocity instead of the rms velocity.) If we assume a sand grain of 52×10^{-4} gm, we get

$$\overline{\overline{V}} = \sqrt{\frac{52 \times 10^{-24}}{52 \times 10^{-4}}} \times (425 \times 10^2) = 10^{-10}(425 \times 10^2)$$
$$= 4.25 \times 10^{-6} \text{ cm/sec}$$

Thus, the rms velocity of our sand grain will be imperceptibly small!

But our proposed task is not quite hopeless if we don't restrict ourselves to the use of pebbles. It is actually possible to make the suspended particles small enough to show noticeable agitation. Thus, if the particles are below about 10^{-4} cm, a lively motion, the **Brownian motion** mentioned in Chap. 19, is observed. (Bacteria can be seen under the microscope to execute Brownian motion.) The smaller the size of the particles the livelier is their motion and the greater is the osmotic pressure, which we obtain with practically realizable concentrations. Very lively motion is rather difficult to observe because of the smallness of the particles which exhibit it. Special optical means (ultramicroscope, see Chap. 30) are necessary for such observations.

Einstein succeeded in deriving an expression for the displacement of a suspended particle from its initial position in a given time. By his formula it is possible to determine Avogadro's number from observations of Brownian motion. The efforts of Einstein, Smoluchowski, and Perrin have made it possible to determine N from observations of Brownian motion by three *independent* methods which yielded identical results. Such quantitative deductions from the kinetic molecular theory and their experimental verification constitute very convincing evidence, supporting our confidence in the kinetic molecular theory.

The osmotic pressure of tiny suspended particles in the so-called *colloidal* suspensions or emulsions is of great biological importance. For example, the wall of our kidney blood vessels is permeable to water and to salts but not permeable to the large protein molecules which, in comparison to, say, sugar molecules, can be considered as a suspension of relatively coarse particles ("colloidal solution"). They give rise to an osmotic pressure of about 4 cm Hg, whereas salts do not exert an osmotic pressure on the blood-vessel wall, since it is permeable to their molecules. As a result there is a tendency for water to filter into the circulatory system owing to the colloid osmotic pressure of the blood proteins. This tendency is counteracted by the blood pressure, which tends to filter protein-free plasma fluid through the semipermeable membrane (which blocks the protein molecules only). Such a filtration process is supposed to be responsible, for instance, for the formation of urine. Filtration can take place only as long as the blood pressure exceeds the colloid osmotic pressure. When blood pressure is lowered to the value of about 4 cm Hg (which may happen in shock), secretion of urine stops.

QUESTIONS AND PROBLEMS

1. Can you explain the mechanism of internal friction (viscosity) in gases?
2. Can you explain why viscosity of a gas may be independent of pressure?
3. What is the probability of throwing a 6 with a die?
4. What is the probability of scoring 12 by throwing two dice simultaneously?
5. Why is the rate of chemical reaction so greatly changed by small temperature changes?
6. How would the pressure of a gas be affected by a change in the velocity distribution of the molecules at constant temperature?
7. Is "Avogadro's law" a law of nature or a hypothesis?
8. At a certain temperature the rms velocity of oxygen molecules is $\bar{v} = 1{,}000$ m/sec. What is the rms velocity (*a*) of helium molecules at the same temperature? (*b*) of a dust particle weighing 10^{-6} gm suspended in air at the same temperature?

9. What property do two different gases at the same temperature have in common?

★10. What kind of molecular velocity is measured in Stern's experiment (Fig. 20.6)?

11. How could you estimate the size of oil molecules?

12. Why is the energy of Brownian motion not dissipated by friction?

13. If a dissolved substance behaves like a gas in exerting osmotic pressure, would you expect this pressure to vary with depth owing to gravity like the atmospheric pressure?

14. How does the osmotic pressure exerted by 1 gm of dissolved NaCl compare with the osmotic pressure exerted by 1 gm of a protein in colloidal solution occupying the same volume?

15. How does the osmotic pressure exerted by 0.1 gm mole of dissolved NaCl compare with the osmotic pressure of 0.1 gm mole of a protein in colloidal solution occupying the same volume?

16. Consider four molecules moving with the following speeds: $v_1 = 100$ ft/sec; $v_2 = 400$ ft/sec; $v_3 = 600$ ft/sec; $v_4 = 1{,}000$ ft/sec. Find (*a*) their average speed, (*b*) their rms speed.

17. How many oxygen molecules, moving at a speed fo 5×10^4 cm/sec, must be reflected from a wall at perpendicular incidence each second in order to produce a pressure of 1,000 dynes/cm^2?

18. A container of 10 cm^3 volume is evacuated to a pressure of 10^{-9} mm Hg, which is the lowest attainable pressure at present. How many molecules does the evacuated tube contain? (Assume $t = 0$°C.)

19. Calculate the rms velocity (*a*) of helium molecules, (*b*) of CO_2 molecules, (1) at −20°C, (2) at 500°C.

20. *a*. How many atoms are there in 54 gm of aluminum?

b. How many molecules are there in 34 gm of ammonia (NH_3)? (See the periodic table of elements in Chap. 36 for the atomic weights.)

CHAPTER 21

ELECTRICITY IN MOTION: 1. MIGRATION OF IONS IN LIQUIDS

a. What are the factors which determine the velocity of migration of particles in electrophoresis?

b. What is an electric current?

c. What is the conventional definition of the direction of an electric current?

d. In what way does the passage of electric currents through electrolytes differ from metallic conduction of current?

e. Are ions which migrate toward an electrode necessarily deposited on it?

f. On what evidence is based the hypothesis of electrolytic dissociation of salts, bases, and acids in aqueous solution?

g. In what part of an electrolysis apparatus does the electrolytic decomposition take place?

h. What were the arguments used by antagonists of Arrhenius' theory of electrolytic dissociation?

i. How do we know that ions are not formed during electrolysis due to the action of the electric field on charged portions of a molecule?

j. Can you state Faraday's first two laws of electrolysis?

k. Can you express the essence of these two laws in one statement?

l. What is one "faraday"?

m. How do you visualize the mechanism of heat generation in an electrolyte traversed by an electric current?

n. What is the source of the electrical energy delivered by an electric battery?

o. What are the essential parts of a voltaic cell?

p. What was the effect of Volta's invention of the voltaic cell on the progress of physical knowledge?

q. How does Nernst's osmotic theory of voltaic cells explain the establishment of a potential difference between the electrodes?

r. What is the "electromotive force"? Does it obey Newton's second law?

s. How much work is done on the charge $Q = 10$ coulombs when it moves once around a circuit which includes a battery of an emf of 20 volts?

t. What is the power supplied by a battery of emf $E = 10$ volts when it delivers a current of 2 amp?

u. Is it possible to construct an electric battery without metallic electrodes?

In previous chapters we studied mechanical and thermal behavior of matter which, in the case of pure compounds, appears like a uniform continuum even under the most powerful microscope. Nevertheless, evidence derived from studies of chemical reactions and of thermal behavior of matter (particularly in the gaseous state) indicated that the assumption of a discontinuous structure of matter would enable us to explain (*i.e.*, correlate) very simply many different phenomena, which otherwise would remain unrelated to each other and to the basic laws of mechanics. The assumption that a chemical compound consists of a large number of tiny identical particles (molecules) is strongly supported by experimental evidence which was presented. The molecules of a compound themselves are assumed to be composed of smaller particles, the atoms, which are supposed to be the smallest portions of a chemical element exhibiting its chemical properties. This has been inferred from the study of chemical reactions between elements.

We are led by these considerations naturally to the question: "Are perhaps other physical entities which we treated so far as continua actually discontinuous, *i.e.*, atomic in structure?" Or more specifically, "Is there any evidence to suggest the hypothesis that the electric fluids, which we treated so far as continua, consist of 'atoms of electricity'?"

In order to answer this question we shall follow a procedure similar to the one which led to the discovery of the discontinuous structure of matter. We shall examine the interaction between electricity and matter and the relation of electricity to various forms of energy in search of observations and ideas, which would enable us to correlate the laws of electricity with laws of mechanics and thermodynamics.

Electric Currents

In some respects the "electric fluids" show a resemblance to gases: They spread over a conductor no matter how large it is, reminding one of a gas expanding into empty space. In one important point, however, they differ conspicuously from gases; namely, the electric fluids at rest reside only on the surface of conductors. Again, just as air will flow from one tire to another one connected to it by a tube until the pressure has become equal in both, electricity will flow from the surface of one conductor to the surface of another one which is in contact with it until both surfaces have assumed the same potential. (As long as the two bodies are at different potential, there is a potential gradient across the conductor joining them; in other words, there is an electric field which moves the charges until the potential difference disappears.)

Let us introduce a term to designate electrical charges in motion. Just as we speak of air currents in reference to moving air, we shall speak of **electric currents** referring to electrical charges in motion regardless of whether they are moving freely or attached to some material objects. *We shall define the strength i of an electric current passing through a conductor as being equal to the charge which passes through the conductor per unit time:*†

$$\boxed{i = \frac{Q}{t}}$$

---

† This defines the average current. The instantaneous current is defined by the quotient $i = \Delta Q/\Delta t$ where Δt is a very short time interval and ΔQ the charge passing through the conductor in that time.

If Q is measured in coulombs and t in seconds, i is given in **amperes.**

We were able to visualize the existence of currents in water by suspending foreign bodies in it. Can we visualize the motion of electric fluids by the same artifice? In some cases we can. We have seen that bodies in contact with a different substance may become electrically charged. Could we not suspend in water small particles, which then would become electrically charged and migrate in an electric field constituting a visually observable electric current?

Electrophoresis

This is possible and the phenomenon is known as **kataphoresis** or **electrophoresis.** For example, red blood corpuscles are electrically charged and will migrate in an electric field. The migration of certain dyes like methylene blue, which yield very fine, charged, suspended particles, can be used for observation of electrophoresis. Kataphoresis of fine, suspended clay or rubber particles has found industrial use in depositing them on a metallic form which is charged so as to attract the suspended particles.

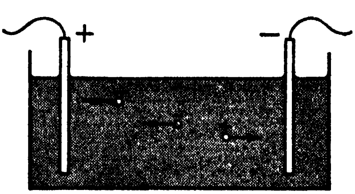


FIG. 21.1. Motion of charged particles in an electric field.

The rate of migration of such charged, suspended particles in an electric field is inconveniently slow for observation owing to their large size in relation to their charge. Exposed to an electric field, as, for instance, that due to the plates + and − of Fig. 21.1 (which are maintained at a constant potential difference), the particles will be initially accelerated but will rapidly reach a constant terminal velocity in the surrounding viscous medium just as rain droplets do as they fall through the air under the influence of the gravitational field.

Electrolysis

Similar phenomena can be observed more easily, owing to the more rapid motion of the particles involved, when solutions of colored salts are used. Figure 21.2 shows a U tube filled at the bottom with a yellowish-red solution of potassium dichromate ($K_2Cr_2O_7$) and with a colorless potassium sulfate solution (K_2SO_4) in the two upper portions R and L of the U-shaped tube. The two solutions form sharp boundaries B_1 and B_2 in the legs R and L of the U tube. E_1 and E_2, the so-called **electrodes,** are metal disks which can be maintained at different electrical potentials, in other words, between which an electric field can be established. The positive electrode is called the *anode*, and the negative one the *cathode*.

As soon as an electric field is established between the electrodes, the boundary B_1 begins to travel upward toward the anode, indicating the presence of some yellowish-red colored, negatively charged particles in our dichromate solution. If, instead of $K_2Cr_2O_7$, we use a blue solution of copper sulfate ($CuSO_4$), we obtain a different result. In this case the boundary B_1 travels downward when the field is established as indicated in Fig. 21.2. This indicates the presence of bluish positive particles in the copper sulfate solution which migrate toward the negative electrode. The speed of migration is very small, being ordinarily in the order of magnitude of a millimeter per hour. We shall call the charged particles which migrate in the above-mentioned experiment **ions** (Greek: wanderers).

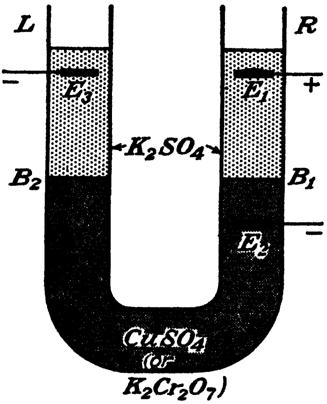


FIG. 21.2. Demonstration of the migration of ions in an electric field.

What is the nature of these ions whose motion constitutes the electric current which flows through the solution between the electrodes E_1 and E_2? The simplest assumption would be to suppose that the negative yellowish-red ions were charged molecules of $K_2Cr_2O_7$ and that the positive blue ions were molecules of $CuSO_4$ which have acquired electric charges through contact with water. How could we test this hypothesis? If this assumption is correct, we should be able, for instance, to precipitate copper sulfate upon the cathode E_2 toward which the positive particles should migrate and where they might be expected to become discharged as they are deposited. This expectation is not borne out by experiment. *We do get a deposit upon E_2, but it is pure copper and not copper sulfate!*

Our solution is evidently decomposed chemically as it passes an electric current. For this reason it is referred to as an **electrolyte**, and the processes of conduction of electricity accompanied by chemical decomposition is called **electrolysis** (*lysis* in Greek means decomposition). It turns out that all acids, bases, and salts in aqueous solution conduct electricity. *These solutions are chemically decomposed in the process of conduction, whereas alloys of metals are not changed chemically even by prolonged passage of very heavy currents.*

Where does the electrolytic decomposition take place? Is the copper split off from the $CuSO_4$ molecule in the electrolysis of copper sulfate as the molecule makes contact with the cathode, or does the $CuSO_4$ molecule dissociate into a positively charged copper portion which migrates toward E_2 and a sulfate (SO_4) portion (which must carry an equal negative charge according to the law of conservation of charge) in the entire space between the electrodes? If positively charged $CuSO_4$ molecules were to exist and to migrate as a whole toward E_2, there should be an increase in the concentration of sulfuric acid (H_2SO_4) at the cathode. But chemical tests show that this is not the case. On the contrary, an increase in the concentration of H_2SO_4 occurs at the anode E_1! This is consistent with the assumption that the $CuSO_4$ molecules split into a positive Cu ion and a negative SO_4 ion in the space between the electrodes. The Cu ions migrate toward the cathode, where they are discharged and deposited. (The process of electroplating is based on such electrolytic deposition of metals.) The negative SO_4 ions migrate toward the anode, where they are discharged and where they immediately react with the surrounding water to form oxygen (which is recognized by the development of gas bubbles) and sulfuric acid:

$$2SO_4 + 2H_2O = 2H_2SO_4 + O_2\uparrow$$

How could one explain the splitting of the copper sulfate molecule which yields the copper and sulfate ions? The following simple hypothesis suggests itself: We can assume that the $CuSO_4$ molecule consists of two electrically charged parts, a positively charged Cu atom (*i.e.*, a copper ion) and a negatively charged group consisting of an atom of sulfur and four oxygen atoms (SO_4). These parts are ordinarily held together by electrostatic attraction. In the electric field between the electrodes E_1 and E_2 the two oppositely charged parts of the molecule should be pulled in opposite directions and the molecule should be torn apart by electrical forces, whereupon the so-formed positive and negative ions would migrate toward the oppositely charged electrodes. This hypothesis sounds plausible. How can we test whether or not it is correct? As usual, we shall make a prediction which follows logically from our assumptions and see whether or not it agrees with experience: If ions are formed by rupture of molecules in the electric field, there should be a minimum force required to tear the ions apart and, hence, a minimum field intensity below which no current could be passed through the solution. Experiments show that such a limiting field strength does not exist. As long as we maintain an ever so small electric field in an electrolytic solution, ions will move in it; *i.e.*, a current will flow.

This experiment leads us to the conclusion that a zero force is required to *ionize* a molecule of $CuSO_4$; in other words, the molecules behave as if they were (at least partly) ionized spontaneously. How can we check the correctness of this conclusion? Let us recall van't Hoff's law of osmotic pressure: According to it, 1 gm mole of $CuSO_4$, containing $N = 6.06 \times 10^{23}$ molecules (Avogadro's number), dissolved in 22.4 liters of water should exert an osmotic pressure of 1 atmosphere at 0°C. If we reduce the number of molecules of $CuSO_4$ in our volume by dissolving a lesser amount, the osmotic pressure should become proportionately smaller. Experiments show, however, that the osmotic pressure decreases less rapidly with increasing dilution than expected. In fact, *at extreme dilutions the osmotic pressure is precisely twice as large as computed for the number of* $CuSO_4$ *molecules dissolved. In other words, there seem to be twice as many particles in our solution as there are* $CuSO_4$ *molecules!*

Dissociation of Electrolytes

This paradox was cleared up by Arrhenius, who worked out the theory of dissociation of electrolytic solutions. He assumed that the molecules of $CuSO_4$ dissociate *spontaneously* when dissolved in water into Cu ions and SO_4 ions. There is supposed to be a state of dynamic equilibrium between the ions and the undissociated molecules. Owing to the molecular agitation two oppositely charged ions may collide and unite to form a $CuSO_4$ molecule, and, conversely, molecular agitation may separate a molecule into its constituent ions. At a given concentration and temperature the ratio of the number of ions to the number of undissociated molecules is constant. Dilution decreases the frequency of collisions between the ions, and hence, the rate of formation of undissociated molecules decreases. The equilibrium shifts in favor of ions, so that at extreme dilutions *only ions* are present. This hypothesis enables us to explain why we find twice the anticipated osmotic pressure for extremely dilute $CuSO_4$ solutions.

Why are the electrolytes which we commonly encounter almost exclusively aqueous solutions? The conspicuously high dielectric constant of water offers a plausible reason for this: The coulomb attraction between the positive and the negative ions should be greatly reduced in a medium of such an unusually high dielectric constant ($K_{H_2O} = 81$).†
This accounts for the high degree of ionization encountered in aqueous solutions.

The hypothesis of electrolytic dissociation and conduction was extended by Arrhenius to all salts, acids, and bases, and its predictions were well confirmed by experiment. It is hard to accept this generalization without asking the question: "But how about compounds like NaCl? We know that free sodium particles react violently with water according to the reaction $2Na + 2H_2O = 2NaOH + H_2\uparrow$." If we are to maintain our hypothesis, we are compelled to assume that the chemical properties of electrically charged sodium atoms (*i.e.* sodium ions) are different from those of (neutral) sodium atoms. While Na atoms do react (in fact violently) with water, Na ions do not. We shall be able to explain the chemical inertness of ions on the basis of Bohr's atomic theory (Chap. 36).

---

† It is noteworthy that water, which is an essential ingredient of all living organisms, is outstanding among all known substances by unusually high values of certain physical constants and by unusual behavior in some respects. For instance, its heat of evaporation, heat of fusion, specific heat, and dielectric constant are extremly high, and its thermal expansion follows an unusual pattern.

Electrolytic Conduction

There is a curious fact connected with electrolytic conduction which calls for an explanation. If we use in the experiment illustrated in Fig. 21.2 a cathode E_3 instead of E_2, we will observe again a downward motion of the boundary B_1 of the copper sulfate solution and an upward motion of the boundary B_2. Evidently all the copper ions in the bent tube travel away from E_1 toward E_3. If the center portion of our U tube were made of rubber tubing several feet long, we could wind it in the fashion of a coil and still observe the same motion of the boundaries B_1 and B_2. What kind of field could make the ions move in circles in the coiled conduit? Should we not expect the ions simply to follow the electric lines of force which for the positive electrode E_1 and negative electrode E_3 should resemble the field diagram shown in Fig. 14.6*b*? Let us examine this question in relation to Fig. 21.2. As soon as the potential difference between E_1 and E_3 is applied, the ions undoubtedly begin to move along the anticipated pattern of lines of force. But as the positive ions moving in the direction from E_1 to E_3 hit the left side of the wall of the tube R, they are stopped and begin to accumulate. Similarly, the negative ions moving from E_3 toward E_1 are stopped at the right side of tube L where they begin to accumulate. These accumulations of charge continue until the established *wall charges* prevent, by electrostatic repulsion, further deposition of charges on the walls. In other words, the initial field between E_1 and E_2 is distorted by the superimposed field of the wall charges in such a way that the field component perpendicular to the wall is zero everywhere in the tube. Thus the only component of the electric-field intensity which can persist is parallel to the tube. Thus the ions move along the axis of the tube no matter how we bend and twist it. (The same question can be raised in relation to current flow in a twisted wire, and the explanation will be analogous to the above case, except that, as we shall see, the charges moving inside a metal are not ions.)

Let us visualize the motion of an individual ion as it is being dragged by electric forces through the water. It should suffer frequent collisions with water molecules to which it will lose part of the kinetic energy which it gained in falling through the potential gradient of the electric field between two successive collisions. In a constant field the ions reach a terminal velocity at which their average energy loss per second due to collisions becomes equal to the energy gain per second due to the motion down the potential gradient (a condition reminding us of the fall of tiny water droplets in air). As the water molecules gain kinetic energy from the ions through the process of collision, the solution gets warmer.

When the ions reach the electrodes, they are discharged and are either deposited like copper in the preceding example or liberated in the form of a gas (for instance, hydrogen is liberated at the cathode and chlorine at the anode when a current is passed through HCl) or they undergo a "secondary reaction" with water such as sodium, which is liberated at the cathode when NaCl is electrolyzed and undergoes the following reaction:

$$2Na + 2H_2O = 2NaOH + H_2\uparrow$$

Convention Regarding the Direction of Current

In the above examples we have been speaking of unidirectional electric currents in cases where two groups of oppositely charged ions move in opposite directions. Whenever the positive ions move west, the negative ones move east and vice versa. How can we speak here of a unidirectional electric current? What direction shall we assign to it? This is purely a matter of convention. **It has been agreed to take the direction of motion of the positive charges as the direction of the electric current.**

Thus, whenever we say the current flows north we may mean one of these three statements:

1. Positive charges move north.
2. Negative charges move south. (It is not necessary that positive charges be present.)
3. Negative charges move south, while positive ones move north.

Faraday's Laws of Electrolysis

Faraday investigated the quantitative relationships between the amount of matter transported and deposited in the process of electrolysis and the strength and duration of the electric current. The setup he used consisted of several electrolytic cells (*i.e.*, vessels filled with an electrolyte containing two electrodes) connected "in series," *i.e.*, in such a way that the same charge which traverses one cell in a given time must also pass any of the other cells in the same time (Fig. 21.3). This arrangement enabled Faraday to compare the amounts of various elements deposited on the electrodes by the same charge or liberated as a gas and collected in inverted test tubes (T_1 and T_2). The use of a current meter M (see Chap. 23) enabled him to determine the effect of the current strength upon the amount of substance deposited in a given time.

Faraday found that

1. For any given substance the mass deposited in a fixed time interval is proportional to the current strength: $m \propto i$.

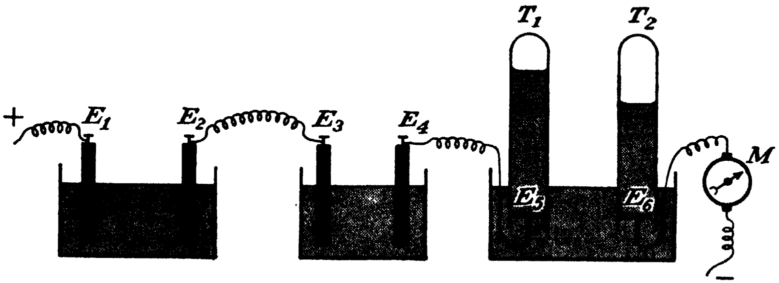


FIG. 21.3.

2. For any given intensity of the current the mass deposited is proportional to the time: $m \propto t$.

3. The amount of a chemical element which is transported along with a unit charge of electricity is proportional to its chemical equivalent: $m/Q \propto A/Z$. (m/Q = mass/charge = mass per unit charge; A = atomic weight; Z = valency; A/Z = chemical equivalent.)

Statements 1 and 2 can be combined into one statement to which we shall refer as **Faraday's first law of electrolysis:** $m \propto it$ or, since $it = Q$, $m \propto Q$

(I) $$\boxed{m = kQ}$$

The amount of matter liberated at an electrode is proportional to the amount of electricity which has passed through the circuit.

The third statement (**Faraday's second law of electrolysis**) can be expressed as follows:

(II) $$\boxed{\frac{m}{Q} = \frac{1}{F}\frac{A}{Z}}$$

The mass transported per unit charge is proportional to the chemical equivalent of the element.

The proportionality constant F is called the **Faraday constant.** Its meaning can be determined as follows:

We can write for expression (II)

$$F = \frac{A/Z}{m} Q \qquad \text{or} \qquad F = \frac{Q}{\dfrac{m}{A/Z}}$$

The ratio $\frac{m}{A/Z}$ is the number of chemical equivalents contained in the mass m. If we deposit precisely 1 gm equivalent of the substance ($m = A/Z$), the denominator $\frac{m}{A/Z}$ becomes equal to 1 and the meaning of F is seen to be the following: **F is the amount of electricity which will deposit one chemical gram equivalent. F has the value of 96,500 coulombs.**

The essence of Faraday's laws of electrolysis may be summarized in the following statement: **It takes the same amount of electricity 1 faraday = F = 96,500 coulombs to liberate by electrolysis 1 gm equivalent of any element** (or complex ion). (One gram equivalent of an ion is defined as its gram molecular weight divided by its valence.)

This law was of great significance in forming our ideas about the nature of electricity. We shall analyze its implications in the next chapter.

Electric Batteries

In our discussion of electrolytic effects observed in solutions in which constant electrical fields are maintained, we have as yet paid no attention to the experimental means which make it possible to establish and to maintain such fields despite the flow of current, which tends to neutralize the charges on the electrodes. The most powerful sources of electrical energy during the eighteenth century were condensers which permitted the maintenance of current of very short duration but were incapable of maintaining a constant potential difference while delivering a current. *New insight in physics is frequently due to new experimental methods which permit the study of phenomena previously inaccessible to observation.* The invention of the electric battery by Volta, which enabled physicists to maintain prolonged electrical currents, led physicists to the acquisition of new knowledge and stimulated the development of new concepts. We can summarize the developments in this field as follows: Galvani, a physician and professor of anatomy in Bologna, discovered what he believed to be "animal electricity." He observed twitching of frog legs when he touched the crural nerve with one metal and a foot muscle with a different metal, the two metals making contact with each other. He believed the "source of the current" to be in the animal tissue. Volta showed that animal tissue in connection with two dissimilar electrodes does form what we now call an "electric battery." But he also showed that animal tissue is not a necessary component of a battery and that **the essential parts of a battery are two dissimilar metals and an electrolyte separating them.** He founded the "contact theory" of electrical batteries according to which the contact of two dissimilar metals was the "pump" driving the electric current around the circuit, the function of the electrolyte being merely to complete the circuit.

Faraday presented the following argument against Volta's "contact theory" of voltaic cells:

> The contact theory assumes that a force which is able to overcome a powerful resistance . . . can arise out of nothing. . . . This would indeed be a creation of power, and is like no other force in nature. We have many processes by which the form of the power is so changed that an apparent conversion of one into the other takes place. . . . But in no case is there a pure creation or a production of power without a corresponding exhaustion of something to supply it.†

Davy showed convincingly that Volta's theory was not correct. He realized that the role of the electrolyte must be more important than merely to close the circuit. The circuit would be a closed one if we were to replace the electrolyte in a battery by a metal, say by mercury, but this combination does not act as a battery. He discovered that voltaic cells (a strip of zinc and a strip of copper in a solution of H_2SO_4) function only with electrolytes which oxidize the zinc. He proposed an alternative explanation of the action of the cell, which, in our modern terminology, is equivalent to the assertion that the electrical energy of the cell is derived from the chemical energy that is released by the chemical processes which take place in the cell while it delivers a current.

Nernst's Theory of Voltaic Cells

We are now confronted with a new form of energy, electrical energy, which seems to be stored in a voltaic cell, and we see as yet no connection between this form of energy and the already familiar forms of heat and mechanical energy. Can we provide a link between the new phenomena and the familiar ones by a suitable hypothesis offering a mechanical picture for the processes which take place in a voltaic cell? Such a hypothesis was offered by Nernst. It proved of great value in suggesting equations for the computation of the potential differences between the

† It is noteworthy that this passage was written by Faraday several years before Helmholtz's publication of his fundamental paper on conservation of energy.

electrodes of voltaic cells. Its theoretical predictions are in good agreement with experimental determinations.

We shall begin with the familiar example of a saturated vapor in contact with the liquid. If equilibrium is established, n molecules leave the liquid surface, evaporating into the surroundings, and just as many molecules condense during this time on the surface coming from the adjacent vapor atmosphere. If, however, the vapor is not saturated, liquid will continue evaporating until the saturation pressure is reached in the space above the liquid surface.

The process of solution is in many ways analogous to evaporation. When a solid, say sugar, dissolves in water, we can imagine sugar molecules to be "evaporating" into the water. As the concentration of sugar increases, the osmotic pressure which its molecules exert upon the solid sugar surface increases; in kinetic molecular terminology this means that the number of impacts of sugar molecules per square centimeter of the solid sugar surface increases.

When will the sugar stop dissolving? This will occur when the number of molecules landing each second on the solid surface is equal to the number of molecules leaving the sugar surface each second, in other words, when the osmotic pressure has become equal to what we shall designate as the "solution pressure," *i.e.*, osmotic pressure of the saturated solution.

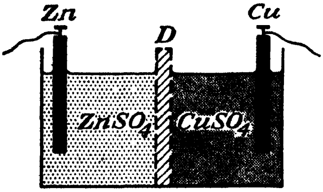


FIG. 21.4. A Daniell cell.

Nernst applied similar considerations to the interaction between metallic electrodes and dissolved ions. Let us consider, for example, a Daniell cell (Fig. 21.4), which consists of a solution of $ZnSO_4$ separated by a porous clay diaphragm D from a solution of $CuSO_4$. The zinc and copper electrodes are immersed in the solutions of their salts.

According to Nernst, a metal surface tends to send positive *ions* into solution when immersed in a solution of one of its salts. This process is analogous to the tendency of, say, sugar to dissolve in water. The metal surface will cease losing ions to the solution as soon as an equilibrium condition is reached where as many ions return from the solution to the metal surface as "evaporate" from the surface into the solution. This case is different from the dissolving of sugar in that the Zn electrode becomes negatively charged with respect to the solution as positive Zn ions leave the zinc plate and go into solution while the solution gains positive ions. Thus, an electric field is set up at the boundary of the electrode which opposes the escape of further Zn ions from the electrode into the solution.

For a given concentration of Zn ions in the solution a definite potential difference between the solution and electrode will be required in order to return to the latter as many ions per second as are leaving it owing to solution pressure. Thus, an equilibrium potential difference is established between electrode and solution whose value depends on the difference between the ionic solution pressure of the metal plate and the osmotic pressure of the ions in solution. According to this picture a metal electrode becomes *negatively* charged when immersed in a solution of its salts when the osmotic pressure of the ions in solution is less than the solution pressure and *positively* charged if the osmotic pressure of the ions exceeds the solution pressure.

In the Daniell cell Zn, which has a high solution pressure, becomes negatively charged with respect to the $ZnSO_4$ solution. The Cu electrode becomes positively charged with respect to the $CuSO_4$ solution, from which we conclude that its solution pressure is relatively low. There is also a potential difference between the two solutions for reasons which are discussed below. Since, however, this potential difference is usually much lower than the potential difference between the Zn and Cu electrodes, we shall neglect it in the following consideration, making, for simplicity, the approximate assumption that the two electrolyte solutions are at the same potential.

We assume that initially some zinc ions go into solution and that some copper ions are deposited from the solution on the copper plate. Once the state of equilibrium has been reached, there is no further deposition of positive Cu ions on the Cu plate because of the repulsive forces exerted on them by the positively charged copper plate. And the zinc plate is not being dissolved any further because the negative charge of the Zn plate prevents further positive Zn ions from leaving the plate. If we now connect the Zn plate to the Cu plate by means of a wire, the opposite charges on both will tend to neutralize each other. A current will flow through the outer wire from Cu to Zn, and since now the charges on both electrodes are less than for an "open circuit," Zn ions can now go in solution and more Cu can be deposited on the Cu plate. The electric field between the electrodes and the respective solutions which previously opposed these processes of solution and deposition has now been weakened. Thus, as long as an external current flows, the Zn plate will continue dissolving and the Cu plate will gain weight. (The reason for using a porous diaphragm *D* is to avoid mixing of the two solutions, and the reason for the use of two different solutions is the desire to avoid electrode polarization.) (See paragraph on polarization.)

If positive electricity were to flow through the wire connecting the two electrodes, its direction of flow would be from Cu to Zn. In the electrolyte positive Zn ions are expelled from the Zn plate into the solution during the process of conduction while positive Cu ions are deposited on the Cu plate. Thus, *positive electricity is moving from* Zn *to* Cu *inside the cell.* In other words, the conventional electric current (motion of actual or of imaginary positive charges) circulates through our system as shown in Fig. 21.5.

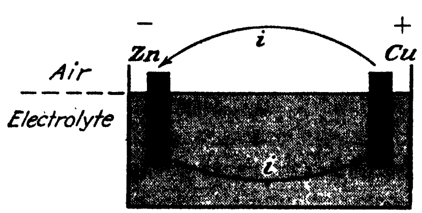


Fig. 21.5.

Electromotive Force

The functions of an electrical battery are similar to the functions of a circulation pump (like, for instance, our heart). Figure 21.6*a* shows a hydrodynamic analogy to the electrical circuit shown in Fig. 21.6*b*. Figure 21.6*c* shows a symbolic diagram of the electric circuit. In the absence of a pump we could impart an impulse to the water in our circulating system of Fig. 21.6*a*. It would circulate for a little while and eventually stop when all the mechanical energy we imparted to the liquid had been transformed into heat by fluid friction. In order to keep the water circulating at a constant rate of *G* gallons per minute, we must continue supplying mechanical energy to our fluid at the rate at which it is being dissipated. Every pound of water must receive each minute as many foot-pounds of energy from the pump as it loses in the same time owing to friction. Thus, we can say that each pound in moving once around the fluid circuit must be compensated for the incurred energy loss by gaining a certain amount of energy, which we might call the "hydromotive work," that must be supplied by the pump.

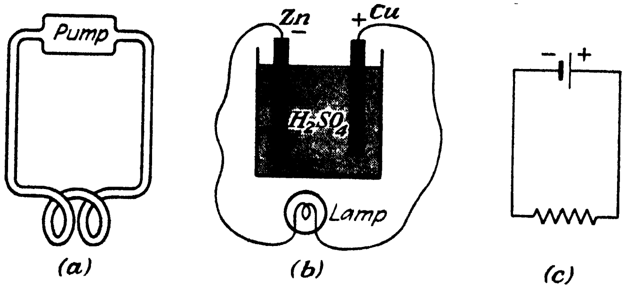


Fig. 21.6. *a*. A closed fluid circulation system and its pump. *b*. A closed battery circuit in schematic representation. *c*. A closed battery circuit in symbolic representation.

Similarly, we have seen that ions lose energy as they move through the viscous solvent. *The role of the battery is similar to the role of the* *pump.* It must supply energy to the moving charges at the same rate at which they lose energy as they move through the solution. In other words, each unit charge of electricity that passes through our "electric pump" (battery) must receive a certain amount of energy in order to maintain a constant current. We could call this *amount of energy which a unit charge receives in moving once around the circuit and which the battery must supply to it the "electromotive work."* Unfortunately, a faulty term has been adopted for this concept. It is called the **electromotive force, or emf.** But we must remember that **it is not a force but work done on a unit charge in moving it once around the electric circuit.**

Power Delivered by a Battery

If we symbolize the emf by the letter E, the amount of energy delivered by the battery to a unit charge of 1 coulomb will be E joules as it moves once completely around the circuit. Hence, the amount of work delivered to the charge Q under these conditions is QE. If the process of driving the charge Q once around the circuit requires t seconds, we can compute the power delivered by the battery by determining the energy delivered per second:

$$\boxed{p = \frac{QE}{t} = E\frac{Q}{t} = Ei} \qquad (21.1)$$

In words: **The power supplied by a battery is equal to its emf multiplied by the current it delivers.** (See pp. 130, 200, and 311 for definitions of units.)

Galvanic Polarization

Two identical metal plates can be made dissimilar by chemical treatment and thus become suitable electrodes for a voltaic cell. Figure 21.7 shows two platinum plates A and C immersed in sulfuric acid: When switch S is in position II, the meter M shows no current flow, since the identical platinum plates immersed in H_2SO_4 do not constitute a battery. When the switch S is in position I, the battery B passes a current through the sulfuric acid, A being the anode and C the cathode. Oxygen is developed at A, and hydrogen at plate C. If we now return the switch S to position II, the meter M will indicate a current, which shows that our cell has become a galvanic battery. This phenomenon is explained by assuming that the platinum plates have become coated with hydrogen and oxygen, respectively; hence, their surfaces are no longer identical. In accordance with Nernst's theory we can ascribe different solution pressures to the modified surfaces. They are said to have been polarized. The same effect (*polarization of the electrodes*) can be accomplished by hanging the platinum plates in atmospheres of oxygen and hydrogen gas, respectively.

This experiment is superficially analogous to the process of charging and discharging a storage battery. In a lead storage battery the current produces a chemical change of the electrode surface. The chemical reaction proceeds in the reversed direction when the battery is discharged. The scheme of charging and discharging the storage cell is the same as that of the polarization cell shown in Fig. 21.7 except that the platinum electrodes are replaced by two lead plates coated with a layer of $PbSO_4$. When the battery is "charged" by passing a current through it, hydrogen is liberated at the cathode and reduces it to pure lead (sulfuric acid being the by-product). At the same time SO_4 ions which lose their charge at the anode initiate the following oxidation reaction:

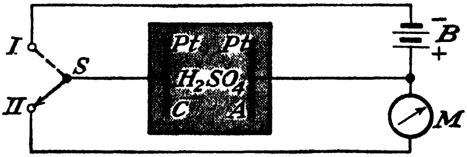


FIG. 21.7. Experiment demonstrating polarization.

$$PbSO_4 + SO_4 + 2H_2O = PbO_2 + 2H_2SO_4$$

The electrode surface thus becomes coated with PbO_2, and we have now two different electrode surfaces, pure lead and lead dioxide. These surfaces immersed in sulfuric acid constitute a galvanic battery, the so-called *charged* **storage battery.** During the process of "discharge" (*i.e.*, when the battery delivers current) the two plate surfaces are transformed back into lead sulfate.

The process of depolarization of electrodes is of great practical importance, especially since the hydrogen liberated and adsorbed on the surface of the anode in batteries such as the "dry cell" lowers the emf of the battery. The contamination of the anode by hydrogen is avoided usually by means of an oxidizing agent. For example, in the case of a "dry cell," which consists of a carbon anode and zinc cathode in a paste of ammonium chloride and zinc chloride which serves as an electrolyte, some manganese dioxide is added to the paste to act as a depolarizing agent by oxidizing the hydrogen liberated at the anode.

In electrobiological work it is important to avoid alteration (by the process of polarization) of an emf applied to stimulate a tissue or picked up from an "active" tissue. This is accomplished by the use of **nonpolarizable electrodes.** Two silver wires on which a layer of silver chloride has been deposited (by reaction with electrolytically liberated chlorine) constitute in contact with tissue fluids (whose electrolytes are mostly chlorides) satisfactory nonpolarizable electrodes. When a current is passed from one of them to the other, the cathode is partly reduced, some of its silver chloride being reduced to silver, whereas the anode is oxidized by formation of more silver chloride. But as long as both electrodes retain silver-silver chloride surfaces, the battery which they constitute when immersed into the tissue chlorides yields a negligible emf.

APPENDIX

Examples from Electrobiology

Aqueous solutions of electrolytes form a vital part of all living organisms. We know that there is a potential difference maintained between the interior of cells and their environment. This potential difference usually undergoes a variation during the activity of an organ which consists of such cells. We don't know precisely what role these electrical processes play in the maintenance of life, but we can at least consider them as revealing electrical symptoms of biological activity, the study of which may lead to new insight into biological processes.

Before presenting biological examples, let us consider how it is possible for potential differences to be set up between different electrolytes in the absence of metal electrodes.

Concentration Potentials

Suppose we pour a solution of fairly concentrated HCl into a test tube and carefully pour a very dilute solution of HCl over it (Fig. 21.8). Both the H and Cl ions will diffuse upward beyond the original line of demarcation *D* between the two solutions, but the lighter H ions (whose average velocity due to molecular agitation is greater than that of Cl ions) diffuse more rapidly. Thus, the front of ions diffusing from space *A* into space *B* is led by a "vanguard" of positive H ions. This vanguard constitutes a positively charged ion cloud. The "rear guard" is a mixture of H and Cl ions which is negatively charged owing to a predominance of the lagging Cl ions. This separation of charges by diffusion tends, however, to check the rapid advance of the H ions. They are retarded due to the attraction by the excess chlorine ions of the lagging negative cloud. Conversely, the chlorine ions in the rear guard cloud are attracted toward space *B* by the hydrogen vanguard. Eventually, an equilibrium condition is reached in which the velocities of diffusion of the H and Cl ions have been equalized by the forces of electrostatic attraction.

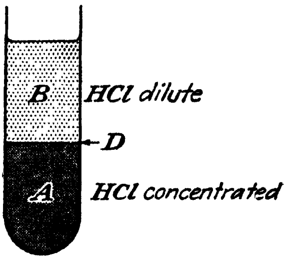


FIG. 21.8. A concentration cell.

The electric field which is set up between the vanguard and the rear guard ions is associated with a gradient of potential, as a result of which a potential difference exists between the lower *A* and upper portion *B* of our solution. The existence of a potential difference between *A* and *B* can be demonstrated by the insertion of two *identical* electrodes, placing one near the surface of *B* and the other near the bottom of *A* and measuring their potential difference. Since the electrodes are identical, the only asymetry in the system which can be responsible for the existing difference in potential is the concentration gradient in the electrolyte.

The situation would be more easily visualized if D were a membrane permeable to H ions but impermeable to Cl ions. (Such membranes, showing selective permeability to ions, can be produced artificially and do occur in living organisms.) Then the positive H ions above D would not be able to penetrate into the space B far above the membrane, owing to retarding forces exerted by the Cl ions which have been left behind. On the other hand Cl ions being attracted by the positive H ion cloud on the opposite side of the membrane will move close to the membrane, which will thus be sandwiched between a positive and a negative layer of ions (an "electric double layer"). An electric field will be maintained across the membrane, and there will be a potential difference between the fluids on the two sides of the membrane.

Electrical Symptoms of Biological Activity

Volta's arguments against the explanation of Galvani's experiments on the basis of the hypothesis of "animal electricity" were so convincing that this hypothesis was generally given up. And yet protagonists of the hypothesis of animal electricity were not entirely wrong. They made a correct assertion which they inferred from facts from which their assertion actually did not follow. Later researches showed that it is possible to detect potential differences in living tissues using two identical electrodes; this excludes the possibility of interpreting these potential differences as being of the same nature as those encountered in voltaic cells, where the presence of two dissimilar metals is essential.

One finds in living organisms potential differences of two kinds: (1) relatively *constant ones,* indicating a permanent state of tissues (for example, the root of a plant is usually at a negative potential with respect to the leaves or with respect to the stem) and (2) *transient changes* in potential difference between two points of the tissue which usually accompany organic activity. For instance, a normally contracting muscle fiber or a nerve fiber conveying a message displays electrical changes of the kind described below.

The objectives of research in electrophysiology are to answer the questions: (1) Where in the tissue does the observed potential difference arise? (2) By what processes is it produced and maintained?

The methods employed in this field of investigation are identical with the methods of physics. One makes a reasonable set of assumptions pertaining to the process under consideration and tests the predictions of the assumed hypotheses. The acceptability of the proposed hypotheses is judged by their fruitfulness in leading to predictions verifiable by experiment. One of the main difficulties in this research is that the complexity of biological systems and the lack of precise information on many conditions admit of too many different hypotheses, the relative merits of which are often difficult to appraise. We shall limit ourselves to the presentation of a few facts and of one widely accepted interpretation.

The Potential Gradient at a Cell Boundary

Let us consider, for example, a nerve fiber. The interior of the fiber is at a negative potential with respect to the external environment. How do we know it? We can cut the cylindrical nerve fiber at one end and connect one terminal of an electrometer† (a device measuring potential difference) to the injured spot—which establishes contact with the interior of the cell—and the other terminal to a distant uninjured portion.‡ The injured point is always negative with respect to an uninjured point on the outside of the fiber, regardless of whether the injury is produced by cutting, poisoning, or otherwise damaging the surface of the fiber. As a result, a so-called **injury current** flows from the intact portions of the fiber to the injured ones.

---

† Nowadays vacuum-tube voltmeters are used for such measurements.

‡ Recently, direct determinations of the potential difference across the cell membrane have been made.

How can the potential drop across the normal cell boundary be explained? The following explanation may be accepted tentatively as a plausible one on the basis of present knowledge. The nerve cell manages to maintain a very high concentration of $+K$ ions in its interior (65 times as high as in the surrounding intercellular fluid). At the boundary of the cell we assume the existence of a *cell membrane*. This membrane is impermeable to negative ions, but $+K$ ions can easily pass through it. Now, if the $+K$ ions begin to diffuse out of the cell through the membrane, they are leaving their negative ion partners inside the cell. As described in the preceding paragraph, this leads to the formation of an electrical double layer, with the cloud of positive ions on the outside and of the negative ones on the inside of the cylindrical membrane of the nerve fiber (Fig. 21.9).

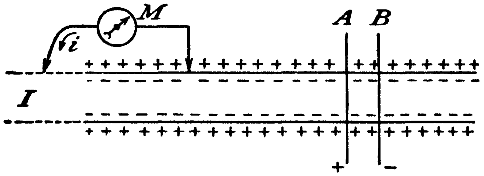


FIG. 21.9. Measurement of potential difference between injured and intact regions of a nerve fiber.

Stimulation of a Nerve

It is very interesting to note that stimulation of the nerve fiber behaves like a temporary *self-repairing* injury. A stimulation can be effected purely mechanically by pressure, by contact with some chemical, or most conveniently (because of the ease of the quantitative control of the intensity of the stimulus) by an electric current.

If two neighboring "stimulating electrodes" A and B are placed along a nerve fiber (Fig. 21.9), a "nerve impulse" (a message) will pass *in both directions* along the fiber as a result of the flow of the stimulating current from A to B. Stimulation can also be accomplished by placing only one electrode on the nerve fiber and the other one elsewhere on the body of the animal. In this case stimulation occurs when the electrode in contact with the nerve fiber is negative; no stimulation occurs when it is positive. In fact, as long as the "stimulating" electrode is maintained at a sufficiently high positive potential, no nerve impulses can pass beyond the region in contact with this electrode. An interesting use of this phenomenon will be discussed below.

Action Potentials of Nerve Fibers

If two electrodes, disposed like A and B in Fig. 21.9, are connected to a quickly responding electrometer, it is noticed that every message passing along the nerve fiber is accompanied by a "wave" of potential variation along the fiber. The potential of an electrode relative to an inactive region varies rapidly, and if we record it as a function of time, we obtain a peak as shown in Fig. 21.10.

This phenomenon can be interpreted as follows: The dotted region of the fiber

(Fig. 21.11) is the stimulated region. It behaves like an injured region in that electrical currents flow from the unstimulated portions of the fiber surface to the stimulated region. The existence and the direction of such local currents are inferred from experimental records such as Fig. 21.10. The state of the stimulated portion of the membrane, resembling injury, can be interpreted as a temporary loss of its selective impermeability to ions. As a result, the double layer is destroyed in that region and its potential drops with reference to the neighboring points on the outside of the fiber. Thus, a potential gradient is established along the outside of the fiber. The outward flow of negative ions takes place along the edge where the stimulated region borders on the unstimulated portion of the fiber, *i.e.*, at the boundary between the solid and the dotted line in the figure. These local ionic edge currents (*A*, *B*, *C*, *D* of Fig. 21.11)† make the region of the "membrane breakdown" (dotted region of Fig. 21.11) grow in both directions (*i.e.*, to the right and to the left), and one would expect that one single stimulus would injure the entire cell membrane and render it ineffective. But there is a spontaneous recovery of the state of selective impermeability of the membrane and of the potential difference across it, so that the point first stimulated is the first point to recover. Thus, as the region of breakdown grows

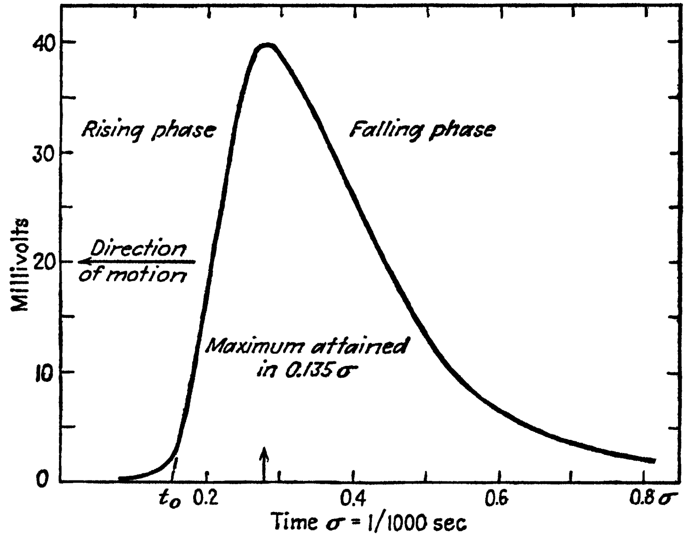


FIG. 21.10. Potential variation at the "active electrode." (Myelinated mammalian nerve fiber.) (*From Stuhlman, "Biophysics," Wiley.*)

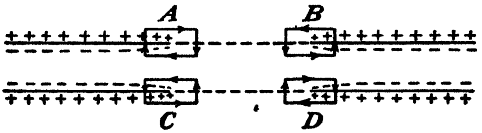


FIG. 21.11.

---

† The arrows in Fig. 21.11 indicate the direction of the conventional positive current.

in both directions, it is followed by a region of recovery, which also issues from the point under the stimulating electrode (which is the first point to recover) and spreads out. The region of local current flow is thus sandwiched between points as yet unstimulated and points recovering from stimulation. The region of local current flow travels in both directions along the fiber. It is the electrical symptom of the biological message transmitted by the nerve.

If we should take at some moment a "snapshot" of the conditions prevailing along a nerve fiber a short time after its stimulation, we should obtain about the picture shown in Fig. 21.12.

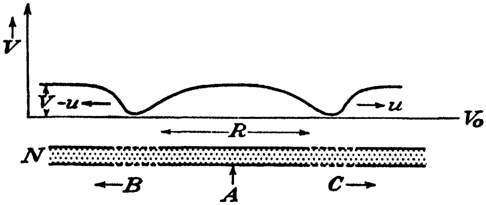


FIG. 21.12. The spreading of a "wave of negative potential" along a nerve fiber (N).

A is the point of original stimulation. R is the region of spreading recovery. B and C are the regions of membrane breakdown and local currents, which are moving away from the point A. The upper portion of the diagram shows the electrical conditions along the fiber surface. V_0 is the base line corresponding to the potential of the interior of the fiber. V is the positive potential of the exterior of the fiber with reference to the potential in its interior. We see that the potential at B and C is negative with respect to the environment and that a *wave of negative potential* (the potential variation which accompanies the region of "membrane breakdown" of B and C) is spreading along the fiber (with speed u ranging from 0.5 to 100 m/sec).

Electroanesthesia

It has been shown that a current flowing into the nerve fiber (a condition which exists at the anode) is capable of blocking the passage of nerve impulses through it. Would it be possible to utilize this effect, which is the inverse of excitation, to render portions of the human body insensitive to pain or even to induce a state of general anesthesia? The advantages of electrical anesthesia are obvious: It would be easy to control the depth of anesthesia; the anesthesia could be easily interrupted or maintained for long periods of time without toxic effects. The advances made in this field are quite encouraging. There are reports of successful amputation of limbs under local electroanesthesia. General electroanesthesia has been maintained in dogs by Silver and Gerard by passing a direct current from an anode in the dog's mouth to a rectal cathode.

Electric Nerve Stimulation

Although a variety of means can be used to stimulate nerves, the method of electrical stimulation is the most convenient one and is most widely accepted chiefly because of the ease in reproducing and gauging the intensity of a stimulus.

Stimulation can be accomplished by a direct current or by an alternating one, *i.e.*, a current the direction of which is changing periodically. The effectiveness of an alternating current in stimulating nerves depends on the frequency of its alternation. The frequency of 60 alternations per second (as used with commercial alternating current distributed for domestic and industrial use) happens to be a most effective frequency for stimulation and, hence, a very dangerous one. As the frequency increases, the stimulating property of the current diminishes and finally vanishes at frequencies about 300,000 cycles/sec. This is one of the reasons why high-frequency currents are used for purposes of internal heating of the human body (diathermy) and for electrosurgery. (Both applications will be discussed in connection with high-frequency currents.)

The process of stimulation appears to be due to a disturbance of the equilibrium of ions which form the electrical "double layers" across the selectively permeable membranes of nerve cells. It is a process vaguely analogous to electrolytic polarization of platinum electrodes discussed on page 324, which produced changes of the electrode surfaces. In the case of nerve cells we can conceive of the nerve fiber as of a potassium electrode, the surface of which may be changed by polarization. The comparison with polarization is helpful in making it plausible why electrical currents become less effective in stimulating nerves as the frequency increases. Let us consider the setup of Fig. 21.7. A unidirectional current liberates hydrogen and oxygen at the cathode and anode, respectively. We can now imagine the direction of the current to be reversed periodically. Thus hydrogen and oxygen will be developed alternately at the same electrode. If the rate of change of polarity is fast enough, the ions of hydrogen (which, when neutralized at low frequencies, would combine with other hydrogen atoms to form molecules of the hydrogen gas) will encounter at high frequencies many oxygen atoms from the previous phase of polarity which have had no time as yet to recombine to form molecules of oxygen gas. Thus, at high frequencies the liberated atoms of H and O will have ample opportunity to combine with each other to form water at each electrode, so that polarization by gas adsorption will completely stop at a sufficiently high frequency. Even though we don't know in detail what processes damage or stimulate the nerve cell through which a current is passing, we are justified by observation of stimulation by currents of different frequencies in assuming that the process is analogous to electrolytic polarization.

Electrical stimulation of nerves and muscles is used as a clinical diagnostic procedure in medicine. It is also used in dentistry to test the "vitality" of a tooth.

Electric Shock and Electrocution

It is an interesting fact that, although in case of **accidental electrocution** by an electric shock (passage of current through the body) the current almost never passes through the brain, the victim loses consciousness. This is considered as a secondary effect caused by the primary stoppage of respiration and circulation.

In most cases of accidental electrocution death is due to passage of the current through the heart. If sufficiently strong, the current will induce ventricular fibrillation; *i.e.*, an irregular wave of contractions, unaccompanied by rhythmical volume changes of the ventricular chambers of the heart, takes the place of the normal heart action. The heart ceases to function as a pump, and circulation stops.

In legal electrocution the current is passed through the brain of the criminal so as to cause immediate loss of consciousness and death by brain damage.

Carefully controlled alternating 60-cycle currents (between 0.3 and 1.2 amp) are passed with the aid of electrodes placed over the temples of mental patients in the so-called **electric-shock therapy**. In this treatment the patient loses consciousness and goes into generalized convulsions.

Electrocardiography

The rhythmical activity of certain organs is accompanied by the generation of electrical emf's which closely parallel the rhythm of the biological activity. An outstanding example of this kind is the heart. It behaves like a battery immersed in a conductive fluid, generating a variable emf. Currents are set up in the entire torso. As a result of these currents potential differences can be detected between different parts of the body. The detection of these **action potentials** of the beating heart is of great clinical importance. They are recorded on a moving film. The potential difference between certain parts of the body as a function of time constitutes the so-called **electrocardiogram.**

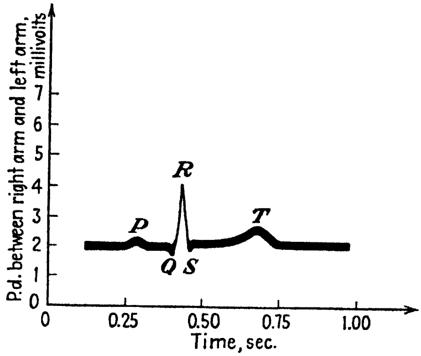


FIG. 21.13. A normal electrocardiogram. (*After Einthoven.*)

Figure 21.13 shows a normal human electrocardiogram depicting the variation in time of the potential difference between the left arm and the right arm. The *P* wave is associated with the auricular contraction, whereas the complex of the *Q*, *R*, *S*, and *T* peaks is associated with the activity of the ventricles.

The study of the electrocardiogram enables the physician to detect among other things cardiac arythmias, myocardial disease, and coronary thrombosis. It is helpful in gauging the effect of the administration of drugs such as digitalis to patients. The significance of deviations of the electrocardiogram record from the normal type is established by generalizations from observations of abnormal records which had been correlated with the accompanying organic symptoms.

Electroencephalogram

Brain cells show a spontaneous rhythmical activity, vaguely resembling the activity of the heart. But in this case, due to the multitude of different cells involved, there is not a single rhythm but rather a melee of different rhythms which range from 1 or 2 per second to about 60 per second (δ, α, and β rhythms). The rhythmical activity of the brain cells is manifested by electrical potential fluctuations of the order of magnitude of 5×10^{-5} volt. These fluctuations of potential difference can be detected by electrodes placed on the skull and recorded by means of a recording electrometer (**electroencephalograph**) which is, in principle, not different from an electrocardiograph.

An electrode placed on the skull is essentially sensitive to electrical changes occurring in its immediate vicinity. The detectibility of the activity in neighboring regions of the brain by the *exploring electrode* diminishes approximately in proportion to the inverse square of the distance of the brain region from the electrode. The other electrode, the so-called *neutral electrode*, is usually attached to an ear lobe. Abnormal conditions, such as brain lesions, are indicated by generation of irregular, slow rhythms (frequency of one-half to two per second). Brain tumors and abscesses can be localized by discovering the source of such low-frequency pulses. This can be done by moving an exploring electrode to different portions of the skull but is practically accomplished more conveniently by distributing 16 electrodes over the skull and connecting them one by one to the electroencephalograph. This method of localization of brain tumors is of great value, because their localization by X rays is generally difficult and often impossible.

Electric Fish

The action potentials generated by the activity of the nerves and of the heart in the human body are very small; they are in the order of magnitude of thousandths of a volt. It is noteworthy that nature has produced animals equipped with special organs for the generation of high electrical potential differences and with the power of discharging intense electric currents through their environment. The outstanding representative of such animals (**electric fish**) is the torpedo which lives in the Orinoco and the Amazon River systems. These fish have poor eyesight and small teeth; they are slow swimmers and are altogether rather helpless, but they are formidable enemies. They are able to stun or render unconscious big animals like horses by a sudden discharge of an electric current. Smaller animals are electrocuted by them and peacefully devoured. They also use their electric power to scare away their enemies by a prolonged series of relatively weak, regular electric pulses. They can develop a maximum potential difference of about 600 volts and can discharge currents of up to about 1 amp. At the same time they are very sensitive to electric currents and can detect potential gradients down to 2×10^{-4} volt/cm. This enables them to communicate with each other by electrical "telegraphy" at rather large distances.

The great physicist Cavendish, though he never saw a torpedo, showed that its stunning power could be explained as being due to an electric discharge. He built an artificial model of an electric fish to prove his point.

Electrophoresis

At the beginning of this chapter we discussed the migration in an electric field of tiny particles which become electrically charged by contact with the fluid in which they are suspended. The fluid, of course, acquires a charge opposite in sign to that of the charge of the particles. The rate of migration of the particles depends on the fluid in which they are suspended, on the temperature, on their size, on their charge, and on the electric-field intensity.

Large protein molecules, which form so-called *colloidal suspensions*, are electrically charged and migrate very slowly in an electric field. In a mixture of proteins the molecules of the different proteins migrate with different speeds in a given field. This fact makes it possible to separate different proteins from each other, which ordinarily cannot be done by chemical methods. An alternative physical method consists of separation by an ultracentrifuge, which is much more elaborate than electrophoresis. This method is of particular value in the separation of serum proteins and the isolation of antibodies in an antiserum.

Electrophoresis can be used to show how an electric current can enter the human body through the skin, which is one of the poorest conductors among the human tissues. This demonstration can be accomplished by using a dye, methylene blue, the particles of which are electrically charged in an aqueous suspension. One puts on the skin a piece of cotton which is wetted by a suspension of methylene blue and places over the cotton an electrode which repels the suspended particles into the skin while another electrode may be applied anywhere else to the body. The electric current, consisting partly of moving, charged methylene blue particles, now flows through the skin. When we take off the cotton and wash the skin, we see a series of tiny blue spots which are not washed off. Microscopic examination shows that they are due to methylene blue particles, which have entered the ducts of the sweat glands. This helps to explain why we are more apt to get an electric shock when we perspire. The moist skin provides a good conducting path along which the current flows to the filled ducts of the sweat glands which conduct it across the skin, by-passing the poorly conductive portions of the skin.

The electrophoretic migration of charged particles into the sweat glands has found an interesting application in the method of **electrical injection of drugs.** The usual method of subcutaneous injection of drugs with a syringe permits the introduction of a limited amount of the drug into a relatively small area from where it is rapidly removed by circulating blood. Certain drugs which acquire an electric charge when suspended or dissolved in water (such as histamine) can be introduced under the skin by electrophoresis (similarly to the "injection" of methylene blue mentioned above). The drug introduced by this method is removed by the blood circulation less rapidly than when injected by needle. Since a large area of the skin can be used for "electroinjection," greater doses of drugs can be introduced and stored subcutaneously, to be given off slowly into the blood stream as if they were secreted by a gland.

Electroosmosis

We can now ask the question: "What happens to the water in the case of electrophoresis? It should acquire a charge opposite to that of the suspended particles; should it not move east when the particles are made to move west in the electric field?" This is a good question. We can show what happens to the water most easily if, instead of using suspended particles of, say, clay, we bind these particles together into a solid structure. In other words, we use a porous clay diaphragm. The walls of the pores in this diaphragm will acquire the same charge as the clay particles in a suspension of clay, and the water will assume the opposite charge. Thus, if we place two wire gauze electrodes between which is maintained a potential difference so as to sandwich the porous clay wall between them, water, which we expect to be electrically charged, is driven constantly through the porous clay wall. This method is used practically for drying peat. Water is driven out of it by means of an electric field maintained across the peat.

The process described above is called electroosmosis. It probably plays an important role in living organisms, although a convincing proof is lacking. Processes like the following are best explained by the assumption of an electroosmotic mechanism:

If a membrane made of the small intestine of a freshly killed rabbit is used to separate two vessels containing NaCl of *equal* concentration, the solution will be pumped across the membrane, moving in such a direction that it leaves the membrane on the side of the "serosa." Since the initial concentration of NaCl is the same on both sides of the membrane, this phenomenon can obviously not be due to osmosis. But it might be due to electroosmosis. (What assumption should we make and what experiment should we perform in order to test whether or not this phenomenon can be explained by electroosmosis?)

QUESTIONS AND PROBLEMS

1. Did we encounter any phenomena up to this point which suggest that the "electric fluids" are discontinuous in structure?

2. Suppose you suspend in water a certain powder whose particles acquire a positive electric charge. Do you expect your suspension to attract small pieces of paper? Why or why not?

3. How could you measure the velocity of electrophoretic migration?

4. Could you devise an electrophoretic procedure (*a*) for separation of different serum proteins? (*b*) for injection of drugs?

5. What is an electric current?

6. What is the conventional definition of the direction of an electric current?

7. Does a charged tennis ball bouncing up and down constitute an electric current?

8. Suppose you have a homogeneous mixture of 10^6 identical positively charged particles (particles A) and 10^6 particles which are identical with the previous ones except that their charge is negative (particles B). What is the direction of the current (*a*) when the A particles move west while the B particles move east? (*b*) when both the A and B particles move east at the same speed? (*c*) when the A particles move east twice as fast as the B particles? (*d*) when the B particles move east twice as fast as the A particles? (*e*) when the A particles stand still while the B particles move west? (*f*) when the A particles move north while the B particles move east?

9. How could you prevent precipitation of fine clay particles on the bottom of a container?

10. How could you deposit fine clay particles on the wall of a cylindrical container without rotating it?

11. The electrodes of a certain electrolytic cell are covered with a coat of insulating paint. Will there be an electric field in the cell when a potential difference is applied to the plates? Will electrolysis take place? Why or why not?

12. Faraday did not have electric batteries which could deliver a constant current through an electrolysis cell for a sufficiently long time. In what way did this constitute a difficulty in establishing his laws of electrolysis? How do you think he overcame this difficulty?

13. Can you think of a situation where the flow of an electric current is not associated with generation of heat?

14. What argument did Faraday use against Volta's contact theory of voltaic cells?

15. What was Davy's contribution to the theory of voltaic cells?

★16. How is the p.d. between the electrodes of a voltaic cell affected by the size of the electrodes? by the amount of electrolyte present? by the concentration of the electrolyte?

★17. Can you think of a conductive link between the poles of a voltaic cell which does not allow any current to flow?

18. How much work is done on the charge $Q = 10$ coulombs in moving it once around an electric circuit which includes a battery of an emf of 20 volts?

19. How do we detect polarization of electrodes?

20. If a battery sends a current through an electrolytic cell whose electrodes are initially not polarized, will subsequent polarization of the electrodes affect the flow of current?

21. What is the power delivered by a battery of emf $E = 10$ volts when it delivers a current of 2 amp?

22. Will the polarization experiment discussed in the text succeed if we replace H_2SO_4 by platinum chloride?

23. Name one mechanism through which potential gradients are established in a living organism.

24. How do you explain the generation of action potentials by active tissues, *e.g.*, nerve fiber?

25. How trustworthy is your explanation? On what evidence is it based?

26. Why does an "electrical impulse" travel along a nerve fiber?

27. Why does the electric eel not electrocute itself? Is it because it is insensitive to the passage of an electric current or because no electric current passes through the eel while it electrocutes its victim?

28. When the nerve of a frog's nerve-muscle preparation is placed on the exposed beating heart of a dog, the frog's leg muscle contracts with every heartbeat. Why?

29. Why does the Daniell cell use two different electrolytes?

30. If you wish to maintain a constant current through a cell consisting of two Pb electrodes in H_2SO_4 solution, is it necessary to apply an emf of a particular value or will any value of emf do?

31. Can you devise a method for measuring electric currents electrochemically? Would such a method be accurate?

32. Can you design a simple portable electrochemical polarity indicator which would enable you to distinguish between the positive and negative poles of a battery?

33. How much copper is deposited on a graphite electrode in a solution of copper sulfate by a current of 3 amp flowing for 10 min?

34. A current of 5 amp flows for 20 min through a solution of sulfuric acid.

a. How many grams of oxygen are liberated at the anode?

b. How many cubic centimeters of oxygen (at standard conditions) are liberated at the anode?

c. How many grams of hydrogen are liberated at the cathode?

35. How many chlorine atoms are liberated at the anode by a current of 10 amp in 2 min? How many oxygen atoms would have been liberated under the same conditions? How many copper atoms? How can you generalize your findings?

36. How many coulombs of electricity must pass through a solution of silver nitrate in order to deposit 1 gm of silver on the cathode?

BIBLIOGRAPHY

Abramson, H. A., and D. A. Moore, On Electrophoresis in Glasser's "Medical Physics," 1st ed., p. 403, Year Book Publishers, Inc., 1944.

Schaefer, "Electrophysiology," 2 vols., 1940–1942.

Gasser, "Electric Signs of Nervous Activity," Philadelphia, 1937.

Coates, Cot, Rosenblith, and Brown, On Electric Fish, *Zoologica*, Vol. 25, p. 249, 1940.

Gerard, R. W., "Unresting Cells," Chap. 8, Harper, 1949.

CHAPTER 22

ELECTRICITY IN MOTION: 2. ATOMICITY OF ELECTRICITY, ELECTRONIC CONDUCTION, ELECTRIC CIRCUIT LAWS

a. What experimental findings suggested the idea that electricity may be atomic in structure?

b. How can we test the hypothesis of the existence of atoms of electricity (elementary charges)?

c. What information enabled physicists to predict the value of the hypothetical elementary electric charge before its direct determination by Millikan?

d. Why is Millikan's experiment more convincing than the method asked for in question *c*?

e. What observation led Millikan to the conclusion that the charge on a body cannot be varied in a continuous fashion?

f. Are elementary positive charges encountered under the same conditions as elementary negative charges?

g. What is the evidence for the assumption that conduction of metals is not due to motion of ions?

h. What evidence suggests to you the assumption that currents in a metal are due to migration of electrons?

i. What is an electron?

j. What is the thermoelectric effect?

k. Why are metals good conductors of heat?

l. What does Ohm's law state?

m. Give an example of a "nonohmic resistance."

n. How do you define resistance of a conductor?

o. What does Joule's law state?

p. How large is the terminal voltage of a battery which supplies a current?

q. When is the terminal voltage equal to the emf of a battery?

The Atomic Hypothesis in Chemistry

In Chap. 21 we did not attempt to interpret a very significant observation. We summarized Faraday's laws of electrolysis by stating that **in electrolytic migration of ions the same amount of electricity, namely, 96,500 coulombs, is transported with 1 gm equivalent of any element.** (One gram equivalent = 1 gm atomic weight over valence, being that amount of a substance which can replace 1.008 gm of hydrogen in a compound or can combine with it.) An analogous observation in regard to matter led chemists to a hypothesis which revolutionized chemical thinking. Chemists noted that 1 gm of hydrogen† will always combine with 8 gm of oxygen to form 9 gm of water and, similarly, that 1 gm of hydrogen will always combine with 16 gm of oxygen to form 17 gm of hydrogen peroxide. Thirty-two grams of sulfur will combine with 32 gm of oxygen (2 × 16 where 16 is the atomic weight) to form 64 gm of sulfur dioxide and with 48 gm (3 × 16) of oxygen to form 80 gm of sulfur trioxide.

---

† More precisely, 1.008 gm.

Attempts to explain why two elements are *not completely* transformed into a compound when mixed in arbitrary mass ratios but form the new compound without a residue only when mixed in certain *discrete* mass proportions have led chemists to the atomic theory of matter. If elements consist of indivisible particles of a definite mass, which varies from element to element, and if formation of a compound involves a definite grouping of these particles, then the laws of definite and multiple proportions appear as a self-evident consequence of the atomic structure of matter.

Atomicity of Electricity

Faraday observed that 96,500 coulombs of *positive* electricity migrate along with 23 gm (*i.e.*, 1 gm atom = 1 gm equivalent) of Na ions in an electrolytic cell, with 39 gm of K ions (1 gm atom), with 32 gm of Cu ions ($\frac{1}{2}$ gm atom = 1 gm equivalent), and with 9 gm of Al ions ($\frac{1}{3}$ gm atom = 1 gm equivalent), and similarly 96,500 coulombs of negative electricity seem to be bound to and thus transported with 19 gm of fluorine ions (1 gm atom = 1 gm equivalent). This association of electrical charges with matter in discrete amounts, so that an integral multiple of 96,500 coulombs is always combined with a gram atom of an element, is closely analogous to the laws of definite and multiple proportions governing the combination of elements among themselves. These laws have been explained by the assumption that elements are made up of atoms. Can we infer from Faraday's observation that electricity, like a chemical element, consists of atoms? Faraday himself recognized the significance of this observation. He writes:

> *Equivalent weights of bodies are simply those quantities of them which contain equal quantities of electricity;* . . . it being the electricity which determines the combining force. Or, if we adopt the atomic theory or phraseology, then the atoms of bodies, which are equivalent to each other in their ordinary chemical action, have equal quantities of electricity naturally associated with them.

Faraday, thus, came very close to the recognition of the atomic nature of electricity. Helmholtz, in a Faraday lecture in 1881, was the first to express clearly the view that Faraday's laws of electrolysis suggest the existence of *atoms of electricity.*

How shall we test whether or not electrical charges are composed of atoms of electricity, which we shall call henceforth "elementary charges"?

One could go about it in the following fashion: (1) Estimate theoretically from available data how large the elementary charge e should be, assuming Helmholtz's hypothesis tentatively to be correct. (2) Design an ultrasensitive electrometer which could detect and measure a fraction of such an elementary charge so that we could see whether or not a body could be charged with a smaller amount of electricity than e units. (3) Attempt to increase the charge on the electrometer in a *continuous fashion* and see if it is possible to do that or if the charge can be changed only jumpwise by discrete amounts, as would be expected on the basis of the atomic hypothesis of electricity. (4) If you find that your electrometer can be charged only with multiple values of an elementary charge e, measure the magnitude of e and compare it with the value predicted in part (1). (5) Finally, investigate whether the atoms of positive electricity are identical with the atoms of negative electricity, (*a*) as to the magnitude of the charge, (*b*) as to their tendency to associate with matter.

Computation of the Elementary Charge. Part (1) of our program can be easily carried out, since we know Avogadro's number (see Chap. 20): $N = 6.023 \times 10^{23}$ molecules/gm mole. Since 1 gm atom of any monovalent element contains 6.023×10^{23} atoms and since the charge which is carried by this number of charged monovalent ions is 96,500 coulombs, each monovalent ion must be associated with the charge

$$\boxed{e = \frac{96{,}500}{6.023 \times 10^{23}} \cong 1.6 \times 10^{-19} \text{ coulomb} = 4.8 \times 10^{-10} \text{ esu}} \qquad (22.1)$$

(See Table I on page 867 for a more accurate value for this charge.) This is, according to our hypothesis, **the smallest charge which can occur in nature.** The charge carried by bivalent ions must be twice as large, by trivalent ions three times as large, and in general, **any charge encountered in nature can be only an integral multiple of this elementary charge** e according to our hypothesis.

This figure holds for the charge of positive as well as of negative ions.

Experimental Determination of the Elementary Charge. The objectives (2), (3), and (4) can be achieved by a proper experimental arrangement, the core of which is a sensitive electroscope. The "oil-drop electrometer" of Fig. 22.1 has been successfully used by Millikan.

Suppose we want to determine the charge q of a tiny, positively charged particle between the plates A and B of Fig. 22.1; how could we do it? It could be done by "electric weighing" of the particle. We could attempt to balance its weight by an upward electrostatic force by making the upper plate negative and choosing the field E between the plates of such a strength that the upward force Eq is equal to the weight mg.

Thus, the equilibrium condition for our particle would be

$$Eq = mg \tag{22.2}$$

In order to find

$$q = \frac{mg}{E} \tag{22.2a}$$

we have to determine the mass m of the particle. $g = 980$ cm/sec^2 is known, and the field strength $E = V/d$ (where V is the potential difference between the plates in electrostatic cgs units and d their separation in centimeters) can be easily measured. The mass m can be found from observations of the terminal velocity of freely falling particles in air if we use *spherical particles* such as oil droplets. According to **Stokes' law** (which can be deduced from the laws of motion) the following relation exists between the mass and the (slow) terminal velocity of our tiny falling sphere:

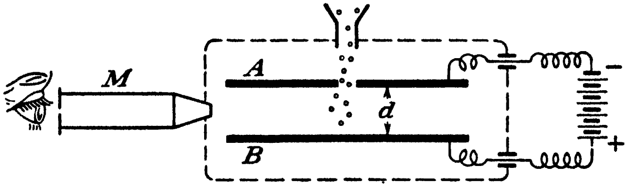


FIG. 22.1. Scheme of Millikan's oil-drop apparatus.

$$m = \frac{4\pi}{3\sqrt{D}}\left(\frac{9}{2g}\eta\right)^{\frac{3}{2}} v^{\frac{3}{2}} \tag{22.3}$$

The acceleration due to gravity g, the density of the oil D, and the viscosity of the air η are known. Hence, from observations of the terminal velocity v of the oil droplets falling between *uncharged* plates, m can be found and substituted in equation (22.2a). The unknown charge q is then determined by substituting in equation (22.2a) the value of the field intensity $E = V/d$, which is ascertained by measuring the value V (of the potential difference between the condenser plates A and B) at which the droplet remains suspended in space. Observations are made by means of a microscope M (Fig. 22.1). The illuminating light is directed at right angles to the plane of the paper.

Millikan was able to watch a single droplet at times for as long as 4 hr. He found that a neutral droplet can only suddenly acquire a charge but never gradually. The sudden acquisition of charge is evidenced by a sudden acceleration of a neutral particle, which falls initially without acceleration with its terminal velocity in the electric field between the plates A and B. The droplets lose their charge just as suddenly as they acquire it, and the sign of the charge of a droplet occasionally changes.

The charge of an oil drop is often seen to increase in magnitude, and like all observed variations of charge, this is a discontinuous process. (The charge of a droplet may change spontaneously, but a more frequent incidence of such changes of charge is secured by irradiation with X rays.)

Millikan found that his oil drops could not acquire a charge less than $e = 1.602 \times 10^{-19}$ coulomb and that the charge of the droplets could change only by integral multiples of this amount. This is an excellent confirmation of our hypothesis.

As to the point (5*a*) of our program of investigation, Millikan's experiment as well as electrolytic observations show that the value of the elementary positive charge is equal to the value of the elementary negative charge.

Part (5*b*) of our program requires a more extensive investigation. In electrolytic experiments positive as well as negative "elementary charges" were found only in association with matter. In order to study the interaction of positive and negative electricity with matter, one would attempt, first of all, to dissociate electricity from matter, in other words to isolate "atoms" of "pure electricity," if possible of positive as well as of negative sign.

Passage of Electricity through Gases and through Empty Space

In order to achieve our objective of isolating electric charges from matter, we shall try to create electric charges in a vacuum. A perfect vacuum cannot be achieved in our laboratories; we can at best obtain a space filled with highly rarefied gas. We shall pursue our aim by studying the passage of electric currents through gases of varying degrees of rarefaction.

If a high enough potential difference is maintained between the electrodes A and C of a tube (Fig. 22.2) which contains a gas (*e.g.*, neon) at a sufficiently low pressure (*e.g.*, $p = 0.01$ mm Hg), a current can be passed through the tube. The flow of current is accompanied by emission of light, which makes such a tube suitable for illumination purposes (*e.g.*, neon signs). The passage of the current can be easily demonstrated by putting an electrolytic cell in series with the tube.† In the cell shown in Fig. 22.2 chlorine is liberated at plate P and copper is deposited at plate N. The gas in our discharge tube is pure neon. Is the conduction of current through it due to migration of neon ions? Toward which of the two electrodes do neon ions migrate? What ions migrate to the opposite electrode?

Experiments show that positive neon ions move toward the cathode C. If we drill a small hole H_1 in the cathode, which is maintained at negative potential with respect to *A*, we can detect a luminous beam issuing from the hole into the space to the left of *C*. Analysis† shows that this luminous beam is a jet of fast-moving positive neon ions which have been accelerated toward *C* and have chanced to pass through the hole H_1. This beam is called a **canal ray.**

---

† In practice one would use an ammeter for this purpose (see Chap. 23).

If we drill a hole H_2 in plate *A*, we can similarly detect a luminous beam passing into the space to the right of *A* presumably due to negatively charged particles which have been accelerated toward *A* and have slipped through the hole H_2. These experiments demonstrate the presence of negative as well as of positively charged particles in our "gas-discharge tube." The motion of these particles constitutes the electric current.

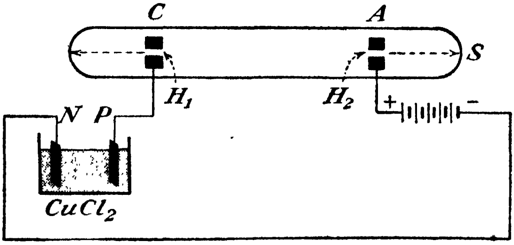


FIG. 22.2.

As we continue to reduce the pressure toward a value of about 10^{-5} mm Hg, we obtain less and less visual evidence of the negative ray in the gas space to the right of *A* whereas the positive canal ray remains visible to the left of plate *C*. Although we cannot detect any fast-moving negative neon ions in the ray entering the space to the right of *A*, there is evidence of a flux of negative charges which pass through H_2, moving in the direction toward the anode *A* (**cathode rays**). The cathode ray is revealed by a green spot created opposite the hole H_2 at the point *S* at which the fast-moving negative charges presumably strike the glass of the tube and make the glass luminous (**fluorescence**). The negative sign of these moving charges can be established by noting which way they are deviated in an electrostatic field.

When the pressure in the tube is reduced to a considerably lower value (say below 10^{-6} mm Hg), both the positive canal ray and the negative cathode ray disappear and the tube ceases to conduct electricity. The tube can, however, be rendered conductive again by heating both electrodes *A* and *C* to incandescence. Under these conditions the positive canal ray (to the left of *C*) does not reappear but the cathode ray (to the right of A) is present again. When C is allowed to cool, the cathode ray disappears, but when only A is allowed to cool, the cathode ray persists. Thus we see that the cathode C, when left at a high temperature, acts as a source of negative electricity which flows from it toward the cold "plate" A, part of it leaking through the hole H_2 to form the observed "cathode-ray spot" at S. At the same time no positive ions are observed in this highly evacuated tube.

---

† Spectroscopic methods of analysis applicable to this study are discussed in Chap. 31.

The Electron Hypothesis

In the above experiment with the incandescent cathode, negative electricity leaves the electrode C, enters the plate A (assume now the holes H_1 and H_2 to be plugged), and moves through the rest of the electric circuit in a continuous cycle. While liberation of Cu and Cl_2 is observed in the electrolytic cell at the plates N and P, no such deposition or liberation of matter of any kind is observed at the plates A and C. This suggests that we may be dealing here with the flow of the "atoms" of *negative* electricity through our vacuum tube. We shall give them tentatively a name: "negative electrons," but for brevity, until we discover their positive counterpart, we shall refer to them as **electrons.**

The search for positive electrons by investigating transport of electricity in liquids, solids, and gases proved wholly unsuccessful. Positive electricity was always found associated with matter, forming positive ions. Only negative electricity could be observed in a free state. (Eventually the positive counterpart of the negative electron was discovered in the cosmic radiation. The properties of this particle will be discussed in Chap. 39.)

To explain various phenomena of electric conduction *we shall adopt tentatively the hypothesis that negative electricity can be considered as a fluid resembling a gas whose smallest indivisible particles (electrons) have a small mass and a charge which is equal to the elementary charge determined by Millikan.*

Electron Theory of Metallic Conduction and of the Thermoelectric and Thermionic Effects

Conduction of electricity through metals appears to be due to migration of electrons without a simultaneous migration of ions. The assumption of the presence of freely movable electrons in metals is supported by the observation of "evaporation" of electrons from hot metal surfaces (see above). The assumption that metallic conduction is exclusively due to motion of electrons is supported by a simple experiment in which we join, say, a copper wire with a zinc wire and pass an electric current through this combination. If the flow of current had involved migration of ions, one would expect to detect diffusion of copper into the zinc or vice versa. Nothing of that sort is observed even after prolonged passage of intense currents. Additional convincing evidence for the existence of free conduction electrons in the interior of metals will be presented in Chap. 23.

It is possible to set up a complete electric circuit, including a source of emf, in which no electrolytic processes occur at any point, so that, according to our hypothesis, the current in this circuit must be entirely due to migration of electrons. An example of a circuit of this kind, a so-called **thermoelectric circuit,** is illustrated in Fig. 22.3.

When two dissimilar metals [such as bismuth (Bi) and antimony (Sb)] are joined to form a closed loop, an electric current can be made to flow through this loop when the two junctions, *H* (*hot junction*) and *C* (*cold junction*), are maintained at different temperatures. This current can be measured by means of the measuring instrument *M*.† Such a combination of two dissimilar metals is called a **thermocouple.** The emf of a thermocouple can be plotted as a function of the temperature difference between *C* and *H*. If *C* is, for instance, kept at the constant temperature of melting ice, this arrangement can be used as an electric thermometer. We can place the junction *H* at various points at which a determination of the temperature is desired. The reading of the meter *M* can be calibrated to indicate the temperature of the junction *H*. The range of such a thermoelectric thermometer extends far beyond the range of a liquid-column thermometer (such as a mercury thermometer), which is limited by the freezing point and the boiling point of the liquid. Thermocouples are particularly suitable for measurements in locations which are not easily accessible. For instance, the "hot junction" *H* can be enclosed in an injection needle for the determination of intramuscular temperatures.

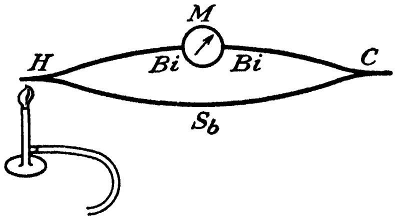


Fig. 22.3. A thermoelectric circuit.

Thermocouples made of very fine wires may be used for measurements of heat radiation. Thus, for instance, heat radiated by the human skin can be picked up by a thermocouple "radiometer" and the temperature of the skin can be measured without placing a solid object in contact with the skin, which would alter the skin temperature.

To offer a mechanical explanation of the thermoelectric effect, we may assume that electrons in the interior of metals move about like molecules in a gas and exert a certain pressure which has different values for different metals at a given temperature. As a result, electrons should flow at the junction H from the metal of higher electron pressure to the metal of lower electron pressure. This would make the junction act as if it were the site of an electron pump, pumping the electrons, say in the clockwise direction, through our circuit. But the identical junction C (*at the same temperature as* H) should pump the electrons at equal rate in the opposite direction. Thus, no electric current should result. But when the temperature of one of the junctions is altered, the equilibrium between the two opposing "electron pumps" is disturbed, because the electron gas pressure at one of the junctions has been increased for one of the two metals more than for the other. The result is a circulation of electrons through the electric circuit, *i.e.*, the thermoelectric current.

---

† See Appendix to Chap. 23 for the principle of such a measuring device.

The electron gas theory of metals is very useful in explaining many other phenomena. For instance, the above-mentioned **evaporation of electrons from metallic surfaces at high temperatures** can be interpreted as follows: At a given temperature electrons inside a metal move about with various speeds. The velocity distribution does not have to be Maxwellian as in the case of gas molecules (and, in fact, it is not). A potential barrier at the metal surface prevents the electrons from escaping; *i.e.*, there is an electrostatic field near the surface which tends to return electrons receding from the metal back to the surface. As the temperature increases, the average kinetic energy of the electrons is augmented and the number of swift electrons goes up. At a sufficiently high temperature of the metal surface an observable number of electrons will possess enough kinetic energy to escape through the potential barrier, giving rise to the so-called **thermionic effect,** which we utilized before to produce cathode rays.

The thermionic effect is used in radio tubes† and X-ray tubes to evaporate electrons from a metal filament which is heated to incandescence by a current passing through it. In an evacuated tube, which contains a cold electrode in addition to a hot filament, electric current can pass only in one direction, namely, so that electrons move from the filament to the cold "plate." This requires that the plate be at a positive potential relative to the filament. A reversal of polarity renders the tube nonconductive. This *valve action* of a two-electrode tube of this type (diode) makes it a valuable *rectifying* device in a-c circuits.

Electron Theory of Metallic Heat Conduction

Metals are characterized as good conductors of heat as well as of electricity. This suggests that the same factor which is responsible for conduction of electricity in metals may also account for conduction of heat. The electron theory represents the process of metallic conduction of heat as follows: The electrons can be assumed to be in thermal equilibrium with the atoms of the metal with which they collide frequently. Due to the small mass of the electrons, their average speed is very high as compared with the average speed of molecules at the same temperature. This enables them to diffuse rapidly in the interior of the metal. When a piece of metal is heated at one end, the increased agitation of the atoms at the hot end is communicated by collisions to the electrons. The increased agitation of electrons is transmitted rapidly by diffusion toward the cooler end, where the molecules begin to gain kinetic energy through collisions with the "warmer" electrons. Thus the very much faster rate of heat conduction of metals as compared with nonmetals is explained on the basis of the same assumption which accounts for their superior ability to conduct electric current, namely, the presence of free electrons.

---

† See Appendix of this Chapter.

Theoretical Formulation of Ohm's Law

We shall finally make an attempt to deduce from our electron hypothesis what kind of a relationship should exist between the potential difference maintained across a metallic conductor and the current passing through it. In the absence of an electric field in the interior of the conductor the motion of the electrons will be a random agitation as pictured previously. If a potential difference is maintained between two points of the conductor, the free electrons will be accelerated in the electric field which has been established in the interior of the conductor. Thus, each electron will gain velocity in the electric field in addition to its instantaneous velocity of random motion. But the gain of velocity in the direction of the field will not continue indefinitely. An electron will be accelerated until it suffers a collision with a metal atom. As a result, it will be retarded and deviated from its original path. Thus instead of constantly gaining kinetic energy by acceleration in the electric field, the electrons will alternately gain it and lose it to the atoms of the metal. An equilibrium state will be reached where the electrons as a whole lose as much energy each second as they gain in the electric field. In this state we can speak of a constant average velocity of electrons, although the velocities of the individual electrons will be quite different and far from constant. The motion of the electrons through a wire in which an electric field exists is analogous to the fall of small rain droplets through air under the influence of gravity. In both cases the loss of potential energy of the particle, instead of resulting in a gain of its own kinetic energy, gives rise to an increase in the kinetic energy of random motion of the molecules of the surrounding substance; *i.e.*, heat is generated. And just as a falling raindrop reaches a constant terminal speed, we can speak of a constant average speed of the drifting electrons under the conditions described above. The terminal speed of the raindrop is proportional to the intensity of the gravitational field, and similarly the average speed of the electron drift can be shown to increase in proportion to the electric-field strength.

We can express the electron current flowing through a wire as we expressed the discharge of water through a pipe:

$$q = dAv$$

(where q is the mass of the fluid passing through a cross section of area A each second, v is the fluid velocity, and d is the density of the fluid). For a stream of electrons we can write

$$i = (ne)Av \tag{22.4}$$

In this equation i is the electric charge transported per second through a cross section of the wire of area A (*i.e.*, the electric *current strength*), n is the number of free electrons† per unit volume (*i.e.*, the electron density), e is the charge of an electron so that (ne) is the freely movable charge per unit volume, v is the average drift velocity of the electrons in the direction of the electric field. If we now assume the average velocity of the electrons v to be proportional to the field strength E, we get

$$v = kE = k\frac{V}{L} \tag{22.5}$$

(where k is a proportionality constant, V the potential difference across the wire, and L its length). Substituting equation (22.5) in equation (22.4), we get

$$i = (ne)Ak\frac{V}{L} \tag{22.6}$$

or

$$i = \frac{V}{L/kAne} = \frac{V}{R} \tag{22.6a}$$

The relation

$$\boxed{i = \frac{V}{R}} \tag{22.6b}$$

is known as **Ohm's law.** It had been established experimentally by Ohm long before the electron theory was proposed.

In the expression $R = L/kAne$, which follows from equation (22.6a), all symbols represent quantities which either are natural constants or are assumed to be constant for a given wire at a fixed temperature. Hence, R must be independent of V and i.

---

† Assumed to remain constant.

The Meaning of Ohm's Law

Equation (22.6*b*) has a double significance: First, it can be considered as a definition equation of a new concept, namely, the concept of **electrical resistance.**

$$R = \frac{V}{i}$$

The ratio of the potential difference between the terminals of a conductor to the current flowing from one terminal to the other is defined as the electrical resistance of the conductor. The unit of resistance is 1 ohm ($1\,\Omega$). A conductor has the resistance of 1 ohm if the potential difference between its terminals is 1 volt when a current of 1 amp passes through it.

The second significant point of equation (22.6*b*) is that it expresses Ohm's law. *Ohm's law* is a genuine law of nature rather than a mere definition equation. It makes a statement which is a generalization of experience. It states:

There is a class of conductors for which the resistance (*i.e.*, the ratio V/i) **is independent of current or voltage across the conductor terminals.**

This is equivalent to the statement that **the voltage across a so-called "ohmic conductor"** (*i.e.*, **one that obeys Ohm's law) is proportional to the current which flows through it** ($V = Ri$). **The proportionality constant R is the resistance of the conductor.**

Not all conductors obey Ohm's law. For instance, this law does not hold for a diode or for a gas-discharge tube, and there is a variety of solid nonohmic conductors. Most of the conductors in practical use, however (like metals and electrolytes), do obey Ohm's law quite closely. Ohm's law is of fundamental importance in the analysis of practical electric circuits.

According to equation (22.7), which follows from equation (22.6*a*), the resistance of a wire should be proportional to its length L and inversely proportional to its cross-sectional area A.

$$\boxed{R = \frac{1}{kne}\frac{L}{A} = \rho\frac{L}{A}} \qquad (22.7)$$

This prediction is found to hold true for ohmic conductors. The meaning of the proportionality constant ρ (the **resistivity**) is apparent when we choose $L = 1$ cm and $A = 1\text{ cm}^2$. Then $R = \rho$; in other words,

ρ is the resistance of a 1-cm cube made of the material in question, measured between two parallel faces of the cube. The value of ρ varies from substance to substance and for a given material usually varies with the temperature. As a rule, ρ increases with temperature for metals and decreases for electrolytic conductors. Certain alloys are known whose resistivity is nearly independent of temperature over a wide temperature range, *e.g.*, constantan and manganin. Wires made of such materials are used as resistance wires in electrical instruments where heating of the wire by the current passing through it must not affect its resistance.†

Joule's Law

A conductor is heated by a current which passes through it. This can be interpreted as being due to the "friction" between the electrons and metal atoms, *i.e.*, due to the fact that the electrons lose the kinetic energy which they acquire in the electric field inside the metal to the metal atoms with which they collide. The rate at which heat is generated in the conductor can be computed as follows: The work W done in moving a charge Q from a point at which the potential is V_1 to a point at potential V_2 is

$$W = Q(V_1 - V_2) = QV \tag{22.8}$$

If this work is done on the charge by the electric field within a time interval of t seconds, the power (*i.e.*, the rate of energy transformation) is given by

$$\underline{P} = \frac{W}{t} = V\frac{Q}{t} = \underline{Vi} \tag{22.8a}$$

By substituting (22.6*b*) in (22.8*a*) we obtain the following equivalent statements of **Joule's law**:

$$\underline{P} = Vi = V\frac{V}{R} = \underline{\frac{V^2}{R}} \tag{22.9}$$

or

$$\underline{P} = Vi = (iR)i = \underline{i^2R} \tag{22.10}$$

These expressions enable us to determine the power dissipated in a given conductor in cases where (1) V and i, (2) V and R, (3) i and R are given.

---

† On the other hand, the variation of the electrical resistance of a wire with temperature can be used for the design of very sensitive and accurate thermometers. So-called **bolometers**, which measure very faint radiations from distant stars, are essentially resistance thermometers. They usually consist of a very thin, blackened platinum foil whose resistance is being measured. The impinging radiation increases the temperature of the foil and thus augments its resistance. The temperature change of the foil can be determined from the change of its resistance. From this temperature change the intensity of the impinging radiation can be inferred.

Ohm's Law for a Complete Circuit

Equation (22.10) should enable us to determine the resistance of a conductor if we know the power dissipated in it as heat by a given current: $R = P/i^2$. Let us consider a circuit consisting of a battery B and an external resistance R_a connected to it (Fig. 22.4)†: When a current flows through this circuit, we notice that heat is generated in the electrolyte of the battery as well as in the external resistance R_a. We can thus ascribe an internal resistance to the battery of the magnitude $R_i = P_i/i^2$, where P_i is the power dissipated in the battery. Similarly, the external resistance is given by $R_a = P_a/i^2$, where P_a is the power dissipated in R_a.

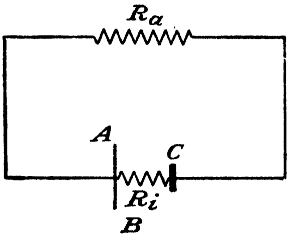


FIG. 22.4.

The total power dissipated into heat in the circuit is

$$\underline{P} = P_a + P_i = i^2R_a + i^2R_i = i^2(R_a + R_i) = i[i(R_a + R_i)] = \underline{i\epsilon} \quad (22.11)$$

According to the definition equation

$$\boxed{\epsilon = i(R_a + R_i)} \quad (22.12)$$

the meaning of the new symbol ϵ introduced in equation (22.11) is that of the sum of the potential drops across the external resistance and the internal resistance R_i of the battery. Setting $iR_i = V_i$ and $iR_a = V_a$ we have

$$\epsilon = i(R_a + R_i) = V_a + V_i \quad (22.12a)$$

According to equation (22.8) the work done in carrying the charge Q through the external resistance is equal to QV_a and through the internal resistance of the battery is equal to QV_i. Thus the sum $QV_a + QV_i$ is the total work W done in moving the charge Q once around the complete circuit consisting of the external resistance R_a and the internal resistance R_i.

By multiplying both sides of equation (22.12a) by Q we get

$$Q\epsilon = QV_a + QV_i = W \quad (22.12b)$$

Thus, we infer that $Q\epsilon$ has the significance of the work done in moving the charge Q once around the circuit. From this follows that ϵ **has the significance of work done in moving a** ***unit charge*** **once around the** circuit. **This is the definition of the emf of a battery** as given in the previous chapter. Thus, we see that $\epsilon = i(R_a + R_i)$ *can be considered as a definition equation of the emf of an electrical power source.*

---

† The internal resistance R_i is indicated in Fig. 22.4 between the battery plates A and C. This is not the customary practice. One usually draws R_i outside the battery.

The dimension of an emf is the same as that of a potential difference (work per unit charge); consequently it is also measured in volts.

Equation (22.12) is analogous to equation $V = iR$, *which is an expression of Ohm's law as applied to a part of a circuit, i.e.*, to the resistance R which forms part of a circuit. **Equation (22.12) is referred to as Ohm's law for a complete circuit.**

If we consider that the wires which connect the battery to the external resistance R_a have a negligible resistance, we shall conclude that the potential difference measured across the battery terminals must be the same as the voltage drop across the resistance R_a.† We designate the potential difference across R_a as the terminal voltage of the battery V^*:

$$V^* = iR_a \tag{22.13}$$

From equation (22.12) follows

$$iR_a = \epsilon - iR_i \tag{22.12c}$$

Substituting equation (22.13) in equation (22.12c) we get

$$\boxed{V^* = \epsilon - iR_i} \tag{22.14}$$

The terminal voltage of a battery is less than the emf by the amount iR_i (*i.e.*, **by the internal voltage drop of the battery).**

Evidently $V^* = \epsilon$ *only if either* $R_i = 0$ *or if* $i = 0$. The latter condition is realized when the battery delivers no current. Thus, the emf of a battery can be determined as the terminal voltage when the battery delivers no current. A device called a **potentiometer** makes such a determination possible. Its principle is treated in Prob. 29.

Example 1

Find the emf of a battery which delivers a current of 2 amp when connected across a resistance $R_a = 1\,\Omega$ and a current of 4 amp when it is short-circuited, *i.e.*, when $R_a = 0$.

Solution. When $R_a = 1\,\Omega$, the following equation holds:

$$\underline{\epsilon} = i(R_a + \underline{R_i}) \tag{1}$$

(The unknowns are underlined.)

---

† The voltage drop across a lead wire is $V_L = R_L i$. When the resistance of the lead wire is $R_L \cong 0$, $V_L \cong 0$. Hence all points joined by a lead wire of negligible resistance must be practically equipotential points.

When $R_a = 0$, the following equation must hold:

$$i_0 = \frac{\epsilon}{R_i} \qquad \text{or} \qquad \underline{R_i = \frac{\epsilon}{i_0}} \tag{2}$$

Substituting 2 in 1 we get

$$\epsilon = i\left(R_a + \frac{\epsilon}{i_0}\right) = iR_a + \epsilon\frac{i}{i_0} \tag{3}$$

Collecting terms containing ϵ on the left side of the equation we obtain

$$\epsilon\left(1 - \frac{i}{i_0}\right) = iR_a \qquad \text{or} \qquad \epsilon = \frac{iR_a}{1 - (i/i_0)}$$

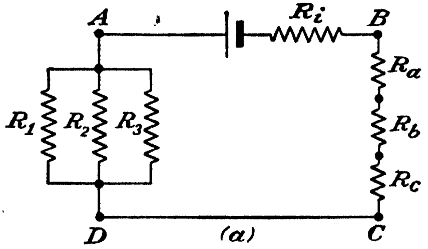
In our case, $i = 2$ amp, $R_a = 1\,\Omega$, $i_0 = 4$ amp.

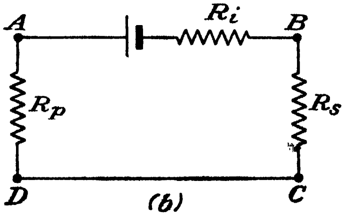
$$\underline{\epsilon} = \frac{2 \times 1}{1 - \frac{2}{4}} = \frac{2}{\frac{1}{2}} = \underline{4 \text{ volts}}$$

$$\underline{R_i} = \frac{\epsilon}{i_0} = \frac{4}{4} = \underline{1\,\Omega}$$

Example 2

In the circuit shown in the diagram the emf of the battery is 4 volts. The internal resistance R_i (drawn outside the battery) is 0.5Ω. The rest of the resistances have the following values: $R_a = 1.5\Omega$; $R_b = 2\Omega$; $R_c = 1\Omega$; $R_1 = 4\Omega$; $R_2 = 2\Omega$; $R_3 = 4\Omega$. Find (*a*) the current delivered by the battery, (*b*) the voltage drop in resistance R_b, (*c*) the current through resistance R_3.





Preliminary Consideration. In order to find the current delivered by the battery, we shall simplify the circuit diagram as shown in Fig. *b*. R_s *is the "equivalent resistance" of the "series combination"* of R_a, R_b, and R_c, in other words, a single resistor across whose terminals a given current will produce the same potential difference as between the points B and C in Fig. *a*. Similarly, R_p *stands for the equivalent resistance of the "parallel combination" of* R_1, R_2, *and* R_3. R_p is an imagined single resistance which will pass the same current for a given potential difference between A and D as the three resistances R_1, R_2, and R_3 do combined. We shall show first how R_s and R_p can be computed.

Computation of R_s. The same current i passes through R_a, R_b, and R_c. According to Ohm's law the voltage drop across these resistors is given by the expressions

$$V_a = iR_a; \qquad V_b = iR_b; \qquad V_c = iR_c \tag{1}$$

The potential difference between the points B and C is equal to the sum of these voltage drops:

$$V_{BC} = V_a + V_b + V_c = i(R_a + R_b + R_c) \tag{2}$$

If we replace the three resistors by the equivalent resistance R_s, we can write

$$V_{BC} = iR_s \tag{3}$$

From equations (2) and (3) follows

$$iR_s = i(R_a + R_b + R_c)$$

and hence,

$$\boxed{R_s = R_a + R_b + R_c}$$

Computation of R_p. Since the lead wires connecting the terminals of R_1, R_2, and R_3 have practically zero resistance, the interconnected terminals are practically at the same potential. Hence, the same potential difference V exists across R_1, R_2, and R_3. According to Ohm's law,

$$i_1 = \frac{V}{R_1}; \qquad i_2 = \frac{V}{R_2}; \qquad i_3 = \frac{V}{R_3} \tag{1}$$

The current i delivered by the battery is equal to the sum of these branch currents:

$$i = i_1 + i_2 + i_3 = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3} = V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right) \tag{2}$$

If we replace these three resistors by the equivalent resistance R_p, we can write for it

$$i = \frac{V}{R_p} \tag{3}$$

Equating (2) and (3) we get

$$\frac{V}{R_p} = V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)$$

Hence,

$$\boxed{\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}}$$

Solution.

$$\underline{R_s} = 1.5 + 2 + 1 = \underline{4.5\Omega} \tag{a}$$

$$\frac{1}{R_p} = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = \frac{4}{4} = 1; \qquad \underline{R_p = 1\Omega}$$

Applying equation (22.12) to the circuit of Fig. *b* we get

$$\epsilon = i[(R_s + R_p) + R_i]$$
$$4 = i[(4.5 + 1) + 0.5]; \qquad \underline{i} = \tfrac{4}{6} = \underline{\tfrac{2}{3}\text{ amp}}$$

According to Ohm's law: $\underline{V_b} = iR_b = \frac{2}{3} \times 2 = \underline{\frac{4}{3}\text{ volts}}$ (*b*)

According to Ohm's law: $\underline{V_{AD}} = iR_p = \frac{2}{3} \times 1 = \underline{\frac{2}{3}\text{ volts}}$ (*c*)

Applying Ohm's law to resistor R_3 we can write

$$V_{AD} = V_3 = i_3R_3$$

or

$$\underline{i_3} = \frac{V_{AD}}{R_3} = \frac{\frac{2}{3}}{4} = \frac{2}{12} = \underline{\frac{1}{6}\text{ amp}}$$

APPENDIX

The Vacuum-tube Voltage Amplifier

Many important advances in modern physics, in physiology, in medical diagnostic techniques, and in technology have been made possible by the invention of the electronic amplifier. The core of an amplifier is a **triode,** *i.e.*, a *vacuum tube* with three electrodes. This tube T is represented schematically as a circle in Fig. 22.5. C is the cathode of the tube which is heated to incandescence by the filament F, which, in turn, is heated by a current supplied by the battery B_f. The hot cathode C gives off electrons which move toward the plate P maintained at a high positive potential with respect to C by the battery B_b. The grid G is an auxiliary control electrode which consists of closely spaced fine wires between which the electrons must pass on their way to P. G is usually kept at a low negative potential with respect to C in order to prevent electrons from landing on it and flowing through the grid circuit.

If the tube contained only two electrodes, C and P, one could vary the voltage V_b of the battery B_b and plot the "plate current" (read on the current meter A) as a function of V_b. In the case where $R = 0$, the plate voltage† $V_p = V_b$ and we obtain a curve shown in Fig. 22.6. The fact that i_p is positive on the left side of 0 means that we still obtain a plate current for small negative values of V_p. This is due to the fact that the electrons evaporate from C with a certain initial velocity and that they can be prevented from reaching P only by an opposing field of a sufficient strength.

---

† V_p is the potential difference between C and P.

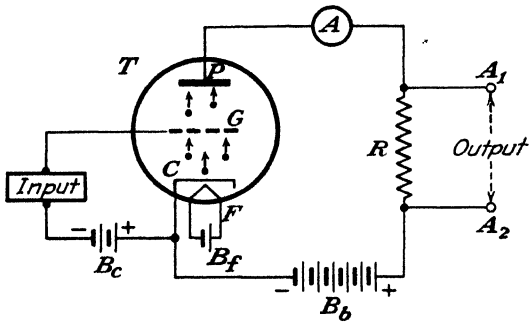


FIG. 22.5. Circuit of a triode amplifier.

From 0 to B, i_p increases with V_p. This is explained as follows: The electrons escaping from C form a negative cloud ("space charge") in front of C. This negative space charge repels the electrons which are leaving C and turns the slowest of them back to C. As we increase V_p, the electrons of the space charge are moved more rapidly. This tends to disperse the electron cloud; its space charge is diminished, and more electrons can reach the plate. Finally, however, we reach a state where an increase in V_p does not increase the plate current ("saturation current"). This happens when all the electrons which are evaporated each second from C reach the plate.

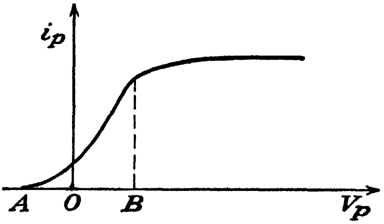


FIG. 22.6. Plate current as a function of triode plate voltage.

The auxiliary electrode G makes it possible to control the electron current which flows to the plate P. Owing to the smaller distance between G and C, the variation in the grid potential V_g required to produce a certain change in i_p is much smaller than the variation in plate voltage V_p which will produce the same change in plate current. The ratio of the changes in plate voltage ΔV_p and grid voltage ΔV_g which produce the same change Δi_p in the plate current (at $R = 0$) is called the **amplification factor** μ of the tube:

$$\mu = \frac{\Delta V_p}{\Delta V_g} = \frac{\text{change in plate voltage}}{\text{change in grid voltage}} \qquad (\text{at constant } i_p)$$

Another important parameter of the tube is the so-called **mutual conductance** g_m. It is defined as

$$g_m = \frac{\Delta i_p}{\Delta V_g} = \frac{\text{change in plate current}}{\text{change in grid voltage}} \qquad (\text{at constant } V_p)$$

Figure 22.7 shows a plot of i_p as a function of V_g taken with the circuit shown in Fig. 22.5 with a "load resistance" $R = 10{,}000\,\Omega$. As the negative value of V_g is diminished, the plate current is increased in a nearly linear fashion if the negative value of V_g is not too high. We see from the graph that, for a change of V_g from -4 to -2 volts, the plate current increases from 3.2 to 4.6 ma.† Thus, we can estimate $\frac{\Delta i_p}{\Delta V_g} = \frac{4.6 - 3.2}{2} = 0.7$ ma/volt $= 0.7 \times 10^{-3}$ amp/volt.

Now, the amplifying property of this circuit depends on the fact that we are able, by a slight variation ΔV_g of the input voltage V_g, to produce a change Δi_p in the current i_p which traverses a high resistance R where this current change produces a change ΔV_R in voltage drop exceeding the variation ΔV_g in the input voltage. Let us suppose, for the sake of concrete illustration, that we vary V_g by $\Delta V_g = 0.1$ volt. How large will be the resulting variation in the voltage drop across R: $\Delta V_R = R\ \Delta i_p$? We know that $\Delta V_g = 0.1$ volt and $\Delta i_p/\Delta V_g = 0.7 \times 10^{-3}$ amp/volt. Hence,

$$\Delta i_p = \frac{\Delta i_p}{\Delta V_g}\Delta V_g = (0.7 \times 10^{-3}) \times 0.1 = 7 \times 10^{-5}\text{ amp}$$

Hence, $\Delta V_R = R\ \Delta i_p = 10^4 \times (7 \times 10^{-5}) = 0.7$ volts

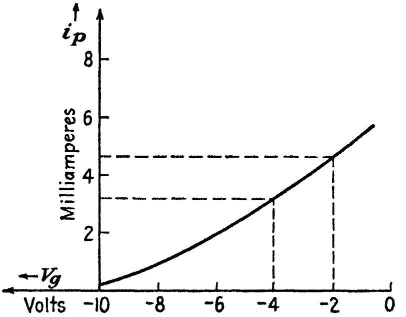


FIG. 22.7. Plate current as a function of triode grid voltage in a circuit containing a plate resistance.

The amplification of our signal is given by the ratio of output voltage variation ΔV_R to the input voltage variation ΔV_g: $A = \Delta V_R/\Delta V_g = 0.7/0.1 = 7$. Thus, in our example, a voltage variation has been amplified sevenfold.

One could now "feed" the fluctuation of the voltage drop across R to the input of a "second stage of amplification" designed similarly to the "first stage" shown in Fig. 22.5 and thus obtain a 7^2-fold amplification of the "signal." By the use of several stages, amplifications in the order of magnitude of millions may be achieved with suitable tubes and circuits.

QUESTIONS AND PROBLEMS

1. Does Millikan's experiment prove that motion of electrons is responsible for the flow of current in a metal?

2. How can we pass an electric current through a vacuum?

3. What is the evidence for the assumption that there are electrons in the interior of metals?

4. How can we ascertain that the thermionic emission is not due to the escape of ions?

5. Can you list the assumptions of the electron theory of metallic conduction?

6. How does the electron theory explain Ohm's law?

7. How does the electron theory explain heating of a conductor by a current?

8. How does the electron theory explain conduction of heat in metals?

9. How does the electron theory explain thermionic emission?

10. How does the electron theory explain the thermoelectric effect?

11. Design a portable thermocouple setup for convenient clinical use to take continuous readings of intramuscular temperatures.

12. Design a simple electric thermometer to study variations in the difference between mouth temperature and blood temperature.

---

† 1 ma = 1 milliampere = 10^{-3} amp.

13. Could you design an instrument more sensitive than a single thermocouple if you had several thermocouples at your disposal?

14. How long is a copper wire 0.1 mm in diameter having a resistance of 1 ohm?

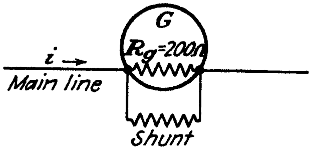
15. A coil is made by winding constantan wire 0.2 mm in diameter on a cylinder 4 cm in diameter and 10 cm long. Assuming the coil to consist of two layers of constantan windings, compute the resistance of the coil.

16. The resistance of a wire at a temperature t_2 can be represented approximately by $R_2 = R_1(1 + \alpha\,\Delta t)$, where R_1 is the resistance at a reference temperature t_1 and α the temperature coefficient of resistance, a constant characteristic of the metal. A platinum coil, used in a resistance thermometer, has the resistance $R_1 = 30\Omega$ at room temperature (20°C). Its resistance is changed to $R_2 = 27\Omega$ when it is submerged in a freezing mixture. Find the temperature $t_2 = t_1 + \Delta t$ of the freezing mixture. (For platinum $\alpha = 0.0038$ per degree centigrade.)

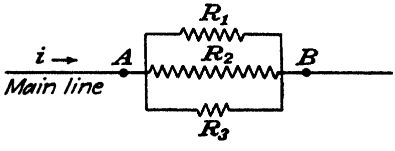
17. A galvanometer (an instrument for measurement of small currents) has a resistance of 200Ω. It gives a full-scale deflection of 100 scale divisions for a current of 10^{-5} amp. Compute the voltage sensitivity of the instrument; *i.e.*, determine the voltage across its terminals which will produce a deflection of one scale division.

18. In order to use the galvanometer of Prob. 17 for measurements of large voltages, how large a resistance should be put in series with it in order to obtain a full-scale reading when 10 volts are applied across the series combination of resistance and galvanometer?

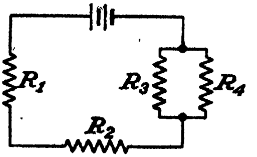
19. The galvanometer G of Prob. 17 can be used to measure heavier currents by connecting it in parallel with a low resistance (shunt). How large a shunt resistance must be connected across the terminals of this galvanometer to obtain a full-scale deflection when a current of 0.01 amp passes through the main line (see figure)?



20. A current of $i = 20$ amp flows through the main line (see sketch). Find (*a*) the potential difference between the points A and B, (*b*) the currents in the resistors R_1, R_2, and R_3. Assume the values $R_1 = 1\Omega$, $R_2 = 2\Omega$, $R_3 = 4\Omega$.

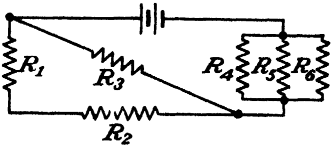


21. In this circuit the emf of the battery is 4 volts, its internal resistance is 0.2 ohm, and the values of the resistances are $R_1 = 1\Omega$, $R_2 = 3\Omega$, $R_3 = 2\Omega$, $R_4 = 4\Omega$.



Find (*a*) the current in R_3, (*b*) the voltage drop across R_4, (*c*) the current in R_1, and (*d*) the voltage drop across R_2.

22. In this circuit the emf of the battery is 6 volts and its internal resistance is 1 ohm. The values of the resistances are $R_1 = 1\,\Omega$, $R_2 = 3\,\Omega$, $R_3 = 4\,\Omega$, $R_4 = 1\,\Omega$, $R_5 = 2\,\Omega$, $R_6 = 3\,\Omega$. Find (*a*) the current in R_3, (*b*) the current in R_5, and (*c*) the voltage drop across R_2.



23. Find the resistance of a lamp rated 40 watts at 110 volts.

24. A lamp rated 40 watts at 110 volts and a second one rated 60 watts at 110 volts are connected in series. If we neglect the variation of resistance with temperature, how much power will the two lamps consume when connected to a battery of 110 volts? How much power is dissipated under these conditions in the first lamp? How much power is dissipated in both lamps when they are connected in parallel to the 110-volt battery?

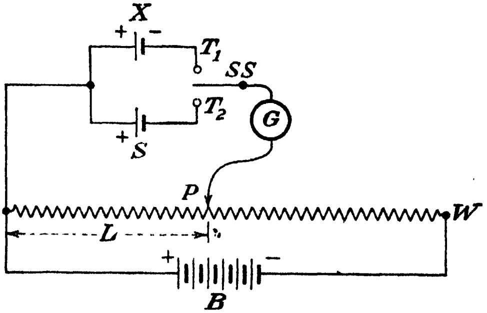
25. How large a resistance must a heater, designed to operate on 110 volts, have in order to heat 1 liter of water from room temperature of 20°C to the boiling point in $\frac{1}{2}$ hr? (Neglect heat losses to the environment.)

26. A motor connected to a 6-volt battery of negligible internal resistance draws a current of 2 amp. What power does this motor consume?

27. How much ice of 0°C is melted by a $10\,\Omega$ heater through which a current of 2 amp is passed for 10 min?

28. In Millikan's apparatus (Fig. 22.1) an oil droplet is suspended at rest between the condenser plates between which a p.d. of 400 volts is maintained, the plate distance being 0.5 cm. Assuming the charge on the droplet to consist of three electrons, compute the weight of the droplet.

29. The diagram shows a **potentiometer circuit** which makes it possible to measure the emf of a cell X, (*i.e.*, its terminal voltage when it delivers no current. S is a "standard cell," *i.e.*, a cell of known emf. The emf of battery B is not known, but

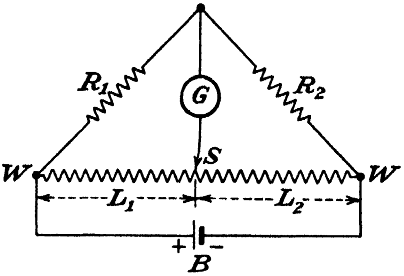


it is constant. *W* is a "slide wire" 1 m long of relatively high resistance on which the contact point *P* slides. The selector switch *SS* makes it possible to connect either the cell S or the cell X to the galvanometer G. According to the position of the contact point P the galvanometer deflects to the left, to the right, or reads zero. A scale makes it possible to determine the length of the wire section L.

a. Explain how this arrangement can be used to measure the emf of cell X.

b. Assume the emf of S to be 1.018 volts and the value of L to be $L = 30.5$ cm when S is connected to the galvanometer. When X is connected to the galvanometer, we find $L = 62.0$ cm. How large is the emf of cell X?

30. The diagram shows a **Wheatstone's bridge.** This is one of the most important electrical circuits. It is of particular interest to us as an example of a group of methods which are classified as *null methods*. It can be made exceedingly precise because (*a*) we can make the slide wire *WW* as long as we please (very much longer than the scale of any pointer instrument) and (*b*) since the galvanometer draws no current when the adjustment is completed, we can use the most sensitive current indicator available. The possibility of using a very sensitive instrument to adjust for zero reading is the general characteristic and main advantage of all null methods. Show that, when the bridge is balanced, *i.e.* when the current through the galvanometer G is zero, the following proportion holds:



$$\frac{R_1}{R_2} = \frac{L_1}{L_2}$$

R_1 is a known resistance. R_2 is an unknown resistance. L_1 and L_2 are the lengths of the two segments into which the sliding contact S divides the wire WW. Assume that in an experiment the following data are given: $L_1 = 37.26$ cm, $L_2 = 62.26$ cm, and $R_1 = 3.596\Omega$. How large is R_2?

31. A dry cell whose emf is $E = 1.48$ volts shows a terminal voltage of $V = 1.42$ volts when a resistance of 3Ω is connected across its terminals. How large is the internal resistance of the cell?

CHAPTER 23

INTERACTION BETWEEN MAGNETIC FIELDS AND MOVING ELECTRIC CHARGES

a. What observations and considerations suggest a relation between electric and magnetic phenomena?

b. What fact did Oersted discover, and how have his findings been generalized?

c. What striking difference is there between the interaction of a magnetic pole with a wire traversed by a current and, say, the interaction of two magnetic poles, two electric charges, or two gravitational masses?

d. How does the force exerted upon a magnetic pole by a straight wire traversed by current vary with the distance of the pole from the wire?

e. What is the origin of the factor $\frac{1}{10}$ in our electromagnetic equations?

f. What is the definition of a "magnetomotive force"?

g. Which way will a wire carrying an east-west current be deviated in a north-south magnetic field?

h. What is an electric current?

i. Is an electric field always required to maintain an electric current?

j. Does a magnetic field always accompany an electric current?

k. Does a falling charged raindrop constitute an electric current? Does it generate a magnetic field?

l. Can you explain the deflection of a wire traversed by a current in a magnetic field on the basis of the electron theory?

m. Does an electron moving at constant speed in a uniform magnetic field in a plane perpendicular to the lines of force experience an acceleration?

n. What is the direction of the force exerted upon an electric charge which moves uniformly through an homogeneous magnetic field in a direction opposite to the direction of the magnetic lines of force?

o. By what hypothesis do you explain the fact that the magnetic flux inside a solenoid changes when an iron core is inserted?

p. By what hypothesis do you explain the fact that we have not succeeded in isolating magnetic north poles from south poles?

q. What experiment can be considered as a confirmation of Ampère's hypothesis of molecular currents?

r. Can you design an arrangement for the determination of the specific charge e/m of electrons? Is it possible to use the same apparatus for ions?

s. What does a mass spectrograph measure?

t. What are isotopes?

u. What methods can be used to separate isotopes?

Relation between Electricity and Magnetism

Certain points of similarity between the phenomena of electricity and magnetism suggested rather early that there might be some relationship between the two phenomena. For instance, both electric and magnetic forces are transmitted through empty space; there are two kinds of poles and two kinds of charges; like poles and like charges repel each other; unlike poles and unlike charges attract each other; both the magnetic as well as the electric properties can be induced in an originally "neutral" body. The suspected relationship between magnetism and electricity was made very probable by the report of an incident in 1731 in which lightning was reported to have struck a building, magnetizing the cutlery through which it passed.

Diverse experiments were conducted in many laboratories in a search for specific interactions between electrified bodies and magnets. For instance, voltaic cells were suspended by fine fibers in the expectation that they might orient themselves in the earth's magnetic field like a compass needle. But nothing of this sort was ever observed.

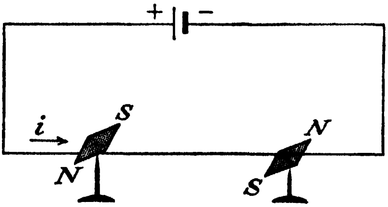


FIG. 23.1. Orientation of magnetic needles in the magnetic field of a current flowing through a wire.

The Magnetic Field of an Electric Current

A crucial discovery was finally made by Oersted, who decided to investigate the interaction between a magnetic needle and a closed electric circuit. Oersted observed that a magnetic needle has a tendency to orient itself at right angles to a wire through which an electric current is flowing (Fig. 23.1). A torque is evidently exerted upon the needle, and this torque must be due to forces exerted in the vicinity of the wire upon the magnetic poles of the needle. We defined a region in which a force is exerted upon a magnetic pole as a magnetic field. Thus Oersted discovered that a magnetic field is set up in the surroundings of an electric current.

What do the magnetic lines of force of this field look like? This can be answered experimentally by spraying iron filings on a horizontal board traversed by a vertical wire through which a current is flowing (Fig. 23.2). The lines of force turn out to be circular and of such a sense that an isolated north pole (N in Fig. 23.2) would be whirled around the wire in a clockwise course looking in the direction of the current.

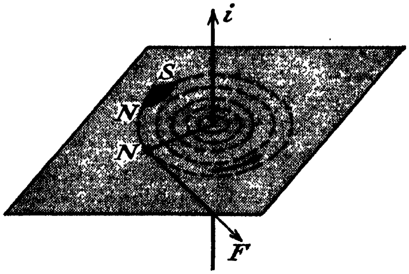


FIG. 23.2. The magnetic field of a wire carrying a current.

The force exerted upon a magnetic pole is seen to be tangential to the circular lines of force. *This observation is a striking one, since in all* *previous cases of two interacting bodies the forces of interaction coincided in direction with a line joining the two bodies. In this case, however, the acting force F is perpendicular to a radius vector drawn from the pole N to the wire.*

The Law of Biot and Savart

How does the magnitude of the force F exerted upon a magnetic pole N depend upon the current, the strength of the pole, and its distance from an infinitely long wire?[†] This question has been answered experimentally by Biot and Savart. The law which they induced is necessarily an idealization from rather crude experiments. Imperfections in their experimental conditions were in part due to the use of a wire of finite length and to the nonexistence of pointlike magnetic poles, which could be localized sharply to determine the field intensity at a point. Rather, they used small magnetic needles (placed near the wire like needle NS in Fig. 23.2) which were allowed to perform oscillations about their equilibrium orientation at various distances from the wire. From the frequency of oscillations (which depends on the orienting torque exerted upon the needles), knowing the pole strength of the needle, its length, and its moment of inertia, they were able to determine the intensity of the local magnetic field at different distances from the wire. They inferred from their data that the *field intensity in the environment of an infinitely long wire* should be given by the expression

$$\boxed{H = \frac{1}{10}\frac{2i}{R}} \qquad (23.1)$$

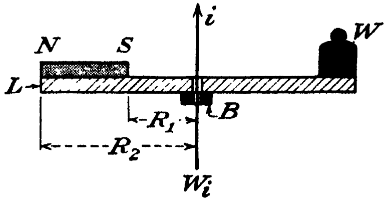


FIG. 23.3.

which is known as the **law of Biot and Savart.** (H is given in oersteds if i is measured in amperes and R in centimeters.)

In view of the crudeness of the method of observation, one might suspect that the relation $H \propto 1/R$ may be only a rough approximation and that perhaps sharper observations would result in an expression containing R raised to a power different from 1, for instance, something like $H = \frac{1}{10}(2i/R^{1.0037})$. This is, however, not the case as was shown by Maxwell using an ingenious and delicate method (Fig. 23.3).

Imagine a stiff, vertical wire Wi through which a current i is flowing. A well-balanced, light, nonmagnetic rod L is mounted on the nearly frictionless bearing B so that it can rotate freely about the vertical wire as axis. We place a magnet NS on one side of L and counterweight W on the other side merely to preserve balance. Owing to the magnetic field of the current, a force will push the pole N toward the reader. According to Biot and Savart's law, this force is given by

---

[†] By this we mean a wire of such a length that the distance from the pole to the wire can be considered as negligible as compared with the length of the wire.

$$F_2 = H_2 p = \frac{1}{10}\frac{2i}{R_2}p \tag{23.2a}$$

(where p is the pole strength of N).

The pole S will experience a force pointing away from the reader of the magnitude:

$$F_1 = -H_1 p = -\frac{1}{10}\frac{2i}{R_1}p \tag{23.2b}$$

The torque due to F_2 with respect to the axis of rotation is

$$T_2 = F_2 R_2 = \frac{1}{10}\frac{2i}{\cancel{R_2}}p\cancel{R_2} = \frac{1}{10}2ip$$

and similarly, the torque due to F_1 is

$$T_1 = F_1 R_1 = -\frac{1}{10}\frac{2i}{\cancel{R_1}}p\cancel{R_1} = -\frac{1}{10}2ip$$

Thus, if Biot and Savart's law is exact, $T_1 + T_2 = 0$ and the arm L should not rotate. If, on the other hand, the power of R is not exactly equal to 1, the sum of the two torques will not cancel and rotation should result owing to the unbalanced torque. A very delicate experimental setup revealed no rotation. Hence, we shall accept equation (23.1) as an exact relationship.

Biot's Law as an Example of a Differential Law. The law of Biot and Savart enables us to compute the field strength at a given distance from a straight, infinitely long wire. But in practice we have to deal with wires of finite length, frequently wound in a complicated way. Thus, this law appears to be of very limited practical utility. It can, however, be used to infer a so-called *differential law*, which could be utilized to compute the field intensity at a given point in the vicinity of a wire of any length or shape. Since the formulation of differential laws is an important general method of physics, we shall illustrate by the example of Biot's law the meaning of a differential law and the usefulness of expressing physical laws in terms of *differential equations*, *i.e.*, equations between differentials (which are small increments in a physical quantity).

Let us imagine the infinitely long wire shown in Fig. 23.4 as being composed of short wire elements ΔL connected in series. Each of these wire elements will give rise to a magnetic field owing to the current flowing through it. The magnetic field at a point P will be the result of summation of all the small contributions to the magnetic field intensity at P made by the individual elements ΔL.

Biot asked himself the following question: How large must be the contribution ΔH made by a small wire element ΔL to the resultant field strength H at P if (as I know) the contributions of all such elements of an infinitely long wire add up to $H = \frac{1}{10}(2i/R)$?

Biot assumed a relationship among ΔH, ΔL, i, R, and θ (see Fig. 23.4) to hold for any wire element and proved by mathematical analysis that the sum of all such expressions set up for every wire element added up to $H = \frac{1}{10}(2i/R)$ for an infinitely long straight wire. His assumed differential equation, known as **Biot's law**, is as follows:

$$\boxed{\Delta H = \frac{1}{10}\frac{i\,\Delta L}{R^2}\sin\theta} \tag{23.3}$$

The proof of the general validity of this differential law is provided by the fact that its mathematical predictions about the field strength at a given point in the environment of a wire of arbitrary length and shape are confirmed by experimental determination.

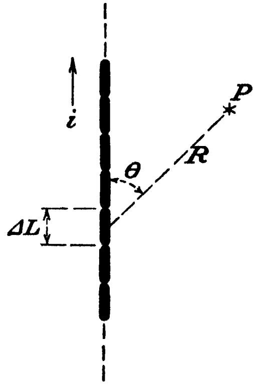


FIG. 23.4.

We shall now consider a few simple examples which show how the field intensity at a given point in the vicinity of a curved wire of finite length can be computed from Biot's law.

The Magnetic Field at the Center of Curvature of a Circular Arc. Let us compute the magnetic field strength at the center of curvature of a circular arc of length s. Inspection of equation (23.3) shows that ΔH has a maximum value in the case where the point P lies along a perpendicular drawn through the center of ΔL. In this case $\theta = 90°$, and $\sin\theta = 1$ has its maximum possible value.

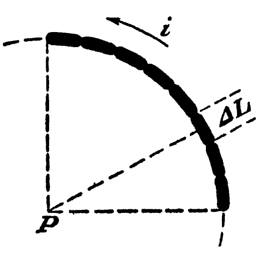


FIG. 23.5.

It can be seen from Fig. 23.5 that lines drawn from the center of curvature P to any of the segments ΔL into which the arc can be divided are perpendicular to ΔL. Hence, the contribution to H at P due to any one of the segments ΔL is

$$\Delta H = \frac{1}{10}\frac{i\,\Delta L}{R^2} \tag{a}$$

The resultant field strength at P is the sum of the contributions of all the segments into which the arc can be subdivided:

$$
\begin{aligned}
H = \sum \Delta H = \sum \frac{1}{10}\frac{i\,\Delta L}{R^2} = \frac{1}{10}\frac{i}{R^2}\sum \Delta L \\
= \frac{1}{10}\frac{i}{R^2}(\Delta L_1 + \Delta L_2 + \Delta L_3 + \cdots) \\
= \frac{1}{10}\frac{i}{R^2}s \qquad (b)
\end{aligned}
$$

In the upper development, i and R have been factored out, since they are the same for all segments ΔL, and $\Sigma\, \Delta L$ has been replaced by the length s of the arc.

Thus we obtain for the **magnetic field at the center of curvature of a circular arc**

$$\boxed{H = \frac{1}{10}\frac{is}{R^2}} \tag{23.4}$$

For a closed circular arc, $s = 2\pi R$; hence,

$$H = \frac{1}{10}\frac{i2\pi R}{R^2} = \frac{1}{10}\frac{2\pi i}{R} \tag{23.5}$$

If, instead of a single loop of wire, we have a *flat* **coil consisting of** N **turns,** we shall get an Nfold field intensity through the summation of the fields which are generated by the individual turns of the coil:

$$\boxed{H = \frac{N}{10}\frac{2\pi i}{R}} \tag{23.6}$$

where H is given in oersteds if i is measured in amperes.

To determine the direction of the magnetic field H at the center of the loop, we merely have to remember that the field is perpendicular to the current and to a line drawn from the center of ΔL to the point P. Thus, we see that *H must be perpendicular to the plane of the loop.*

The Electromagnetic Unit of Current

It is time to remember now that we used the unit of "ampere" without having defined it except in relation to another unit of current: *An ampere is the current which is equivalent to the passage of* 3×10^9 *esu of charge per second.*

The esu of charge was defined in terms of the mechanical interaction of point charges as follows: A point charge of 1 esu repels another identical point charge placed 1 cm from it in a vacuum with the force of 1 dyne.

The force exerted upon a *magnetic* unit pole at the center P of a circular arc (Fig. 23.5) through which an *electric* current is flowing provides an independent method (an *electromagnetic method*) of defining a unit of current. If we do not stipulate that the current be measured in amperes, the factor $\frac{1}{10}$ in equation (23.4) will be replaced by a constant C whose value depends on the choice of the units:

$$H = C\frac{is}{R^2} \tag{23.4a}$$

H is the magnetic field strength or, which is the same, the force in dynes exerted upon a unit magnetic pole at the center of curvature of the arc in Fig. 23.5.

We can now define an **electromagnetic unit of current** (the **abampere**) as follows:

The current through a circular arc of length $s = 1$ cm **whose radius of curvature is** $R = 1$ cm **is equal to** 1 *abampere* **when the force of** 1 **dyne is exerted upon a unit magnetic pole placed at the center of curvature of the arc.**

Under these conditions $H = 1$ oersted, $s = R = 1$ cm, and $i = 1$ abamp by definition, so that our equation (23.4a) becomes $1 = C(1 \times 1/1)$ and, hence, $C = 1$.

It was deemed more convenient to choose a smaller unit of current, **the ampere,** *for practical measurements*: 1 **amp** $= \frac{1}{10}$ **abamp.** With this unit of current C becomes equal to $\frac{1}{10}$.

The electromagnetic unit of charge, 1 *abcoulomb,* follows from the definition of 1 abamp: 1 *abcoulomb is the charge which passes the cross section of a conductor through which* 1 *abamp flows for* 1 *sec.*

Relation between the Electrostatic and Electromagnetic Units of Current

Oersted's discovery provided a link between the phenomena of electricity and magnetism, giving hope of an eventual possibility of unifying the two fields by showing magnetism to be a manifestation of electricity displayed under certain conditions. But at this point, quite unexpectedly, another possibility of eventual correlation with a third seemingly unrelated field of physics appears to be revealed: Experimental determination of a given amount of electricity in esu and at the same time by independent measurements in emu (abcoulomb) established the numerical relationship between the two different systems of units: 1 emu[†] of charge $\equiv$ 1 abcoulomb $= 3 \times 10^{10}$ esu of charge, or

$$\boxed{\frac{1 \text{ emu of charge}}{1 \text{ esu of charge}} = 3 \times 10^{10}}$$

[†] The abbreviation emu stands for electromagnetic units.

The striking feature of this ratio between the values of emu and esu of charge is not the largeness of its numerical value but the fact that it is precisely equal to the velocity of light in a vacuum:

$$\boxed{c = 3 \times 10^{10} \text{ cm/sec}}$$

In fact, this ratio even has the physical dimension of a velocity, as can be shown by dimensional analysis. The discovery of this relationship suggested a close connection between light and electromagnetic phenomena and stimulated one of the most fruitful and imaginative theoretical developments in physics (see Chap. 33).

Magnetic Field as a Manifestation of Moving Electric Charges

Oersted's experiment showed that a magnetic field is generated by an electric current. According to the electron hypothesis the electric current in a metal consists of a migration of tiny electrically charged particles. Can we infer from this that larger electrically charged material bodies will also produce a magnetic field when moved by mechanical means? Would, for instance, a moving charged tennis ball be surrounded by circular magnetic lines of force? This question was answered by Rowland in the affirmative.

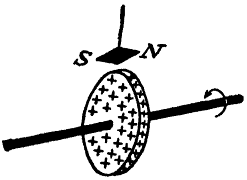
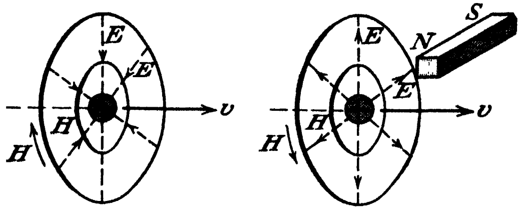


FIG. 23.6. Rowland's experiment.

Rowland performed the following experiment: A charged dielectric disk was rotated as shown in Fig. 23.6. Let us fix our attention on the positive charges located on the rim of the wheel. Their motion is equivalent to the flow of a positive current through a loop of wire which can be imagined to coincide with the rim. We therefore expect that a magnetic needle placed near the rim originally parallel to the plane of the disk will tend to orient itself at right angles to that plane as shown in Fig. 23.6, just as if an electric current were flowing through the rim of the disk in the direction of its rotation. This expectation was confirmed by Rowland's experiment. *Rowland thus showed that the magnetic field is a manifestation of moving electrical charges irrespective of how the charges are moved.* Accordingly, the magnetic field of a current in a straight wire can be considered as the resultant of the magnetic fields of the individual electrons that constitute the current. The drift velocity of the electrons is very slow (of the order of 0.1 mm/sec), but their number is very large, so that the resulting magnetic field strength is appreciable. Figure 23.7 shows the magnetic field of a moving positive charge (*b*) and of a moving negative charge (*a*). Experiment shows that the direction of the magnetic-field lines is opposite in these two cases. On the other hand, if the direction of motion of the plus charges were reversed, the direction of the magnetic field would be reversed, too, resulting in a magnetic field of the same sense as that of the negative charge shown in Fig. 23.7*a*. (The magnetic-field lines are marked by H in the diagram and electric-field lines by E.)



(a) (b)

FIG. 23.7. The magnetic field of moving charges.

Force upon a Current in a Magnetic Field

The arrows of the magnetic-field lines in Fig. 23.7 indicate the direction of the force which a magnetic north pole would experience in the magnetic field of the moving charge. In Fig. 23.7*b* the pole N would tend to move upward. According to Newton's third law, however, the pole N would exert a force of reaction upon the moving charge, which would experience a downward force. **In general, a force is exerted upon an electric charge which moves through a magnetic field, except when the direction of motion coincides with the direction of the magnetic lines of force.** Hence, since the current through a wire is assumed to be due to the motion of a multitude of electrons, we expect a force to be exerted upon a wire which carries a current across a magnetic field. *If the direction of the current i is perpendicular to the direction of the magnetic field H, the force F exerted upon the wire is shown by experiment to be at right angles to both the magnetic field and to the direction of the current* (Fig. 23.8).

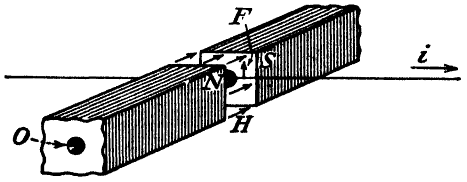


FIG. 23.8. Scheme illustrating the principle of the string galvanometer.

The arrangement of Fig. 23.8 represents the principle of the string galvanometer which is used in "electrocardiographs" to record action potentials of the heart. The current i which is proportional to the potential difference between two points of the human body (which potential difference is caused by the flow of action currents of the heart) is passed through a very fine metallized quartz fiber disposed at right angles to a constant magnetic field which exists in the gap between the pole pieces N and S. The fiber is deviated upward or downward in the magnetic field of Fig. 23.8 in proportion to the strength of the current that flows through it. A channel O drilled through both pole pieces makes it possible to observe or to record photographically the motion of the fiber.

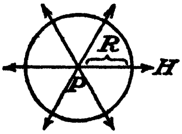


FIG. 23.9.

The magnitude of the force exerted upon the fiber can be inferred from Biot's law and Coulomb's law as follows:

Applying Biot's law to obtain the magnetic field strength at the center of a circular loop (Fig. 23.9), we obtained equation (23.5): $H = \frac{1}{10}(2\pi i/R)$. In this equation H is the force exerted at the center of the loop upon a unit magnetic pole. If instead of a unit pole we placed at the center of the loop a pole of pole strength p, we should get p times as high a force:

$$F = pH = \frac{1}{10}\frac{2\pi i}{R}p \tag{a}$$

We can transform this equation by multiplying the right side by R/R. We get then

$$F = \frac{1}{10}2\pi Ri\frac{p}{R^2} \tag{b}$$

The physical meaning of the term p/R^2 can be interpreted on the basis of Coulomb's law for magnetic poles:

$$F = \frac{1}{\mu}\frac{p_1p_2}{R^2} \tag{c}$$

In this equation F is the force between the poles p_1 and p_2 which are R centimeters apart in a medium of permeability μ. If we make p_2 a unit pole, F will assume the meaning of the force per unit pole or field strength H. Thus for $p_2 = 1$, we get

$$H = \frac{1}{\mu}\frac{p_1}{R^2} \quad \text{or} \quad \frac{p_1}{R^2} = \mu H \tag{d}$$

Considering equations (d) and (b) we see that the term p/R^2 in (b) can be replaced by μH, so that we get

$$F = \tfrac{1}{10}2\pi Ri\mu H \tag{e}$$

Now $2\pi R = L$ is the length of the loop of wire. We can thus write for equation (*e*)

$$F = \tfrac{1}{10}Li\mu H \qquad (f)$$

This is still the expression for the magnitude of the force F exerted upon the pole P at the center of the loop. *But according to Newton's third law this must be also equal to the force which the pole P exerts upon the loop.* Equation (*f*) expresses this force in terms of the length of the loop, the current, and the magnetic field strength H which the pole creates at the distance R at which the loop is located.

Since the magnetic lines of force in Fig. 23.9 are perpendicular to the wire at any point, we should expect the force upon the wire to be the same as in the case illustrated in Fig. 23.8, where the current flows through a straight wire at right angles to a magnetic field. Thus we adopt the same expression for the **force exerted upon a straight wire in a perpendicular magnetic field:**

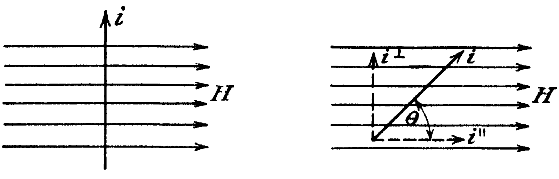


FIG. 23.10.

$$\boxed{F = \tfrac{1}{10}\mu HLi} \qquad (23.7)$$

In this expression H is the field intensity in oersteds, L the length in centimeters of the wire upon which the force F measured in dynes is exerted, i the current strength in amperes, and μ the permeability of the medium which surrounds the wire and magnet. $\mu = 1$ for a vacuum and very nearly so for all "nonmagnetic" materials. This equation is confirmed by experiments with straight wires.

If the wire is not perpendicular to $\vec{H}$, expression (23.7) is still valid if i is taken to mean the component of the current perpendicular to the magnetic field. For example, in Fig. 23.10*a* the force is perpendicular to the plane of the paper and is given by equation (23.7). In Fig. 23.10*b* the force has the same direction as in Fig. 23.10*a* and its magnitude is obtained by resolving the vectors representing the current i into two components: $i^{\parallel}$ parallel to the magnetic field and $i^{\perp}$ perpendicular to it. The $i^{\parallel}$ component is entirely ineffective in generating a force. The force produced is exclusively due to the $i^{\perp}$ component, which is obtained simply from the magnitude of i and from the angle θ, which determines the deviation of i from the direction of H: $i^{\perp} = i \sin \theta$. Hence, we can write a more general expression for F:

$$\boxed{F = \tfrac{1}{10}\mu H L i \sin \theta} \qquad (23.7a)$$

The direction of the electromagnetic force can be conveniently found by the following rule of thumb:

Imagine yourself looking in the direction of $i^{\perp}$ and draw the circular magnetic-field lines associated with $i^{\perp}$. The force will be directed from the region in which these lines reinforce the original field H (*i.e.*, have the same direction as H) **to the region where they oppose the original field** (Fig. 23.11). (The cross section of the wire carrying $i^{\perp}$ into the paper is represented by ⊛ in Fig. 23.11).

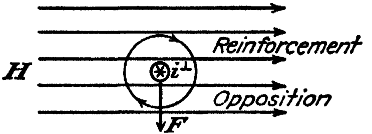


FIG. 23.11. Diagram for the determination of the direction of the force exerted upon a current in a magnetic field.

It is customary to write equation (23.7a) in terms of the vector $\overrightarrow{B} = \mu\overrightarrow{H}$. If, in addition, we assume the current to be measured in abamperes, we shall write

$$\boxed{F = BLi \sin \theta} \qquad (23.7b)$$

The physical significance of $\overrightarrow{B}$ follows from the following consideration. In determining the distribution of magnetic field intensity we place a magnetic test pole at different points of the field and obtain the magnetic-field vector as $\overrightarrow{H} = \overrightarrow{F}/p$, where $\overrightarrow{F}$ is the force exerted upon the pole and p the pole strength. $\overrightarrow{H}$ has the meaning of a force exerted upon a unit pole.

We could similarly explore a magnetic field using a "test current element" instead of a test pole. We could measure the force exerted upon a current element of length L traversed by a current i and oriented so as to experience the maximum force in the field (*i.e.*, oriented at right angles to the field). Then since $\theta = 90°$, we would obtain from (23.7b) for the *magnitude* of B

$$B = \frac{F}{Li} \qquad (23.7c)$$

Setting $L = 1$ cm and $i = 1$ abamp we see that B *has the magnitude of the maximum force exerted upon a "unit current element," i.e., an element* *of unit length traversed by a unit current.* The vector $\vec{B}$ is called the **magnetic induction.** Its direction is perpendicular to the direction of the current and of the force exerted upon the current.

Motion of Free Electrons and Ions in a Magnetic Field

The force which is exerted upon a wire in a magnetic field must obviously be the resultant of all the forces which are exerted upon the individual moving electrons. In Fig. 23.11 the electrons would tend to escape from the wire in the direction of the arrow of F but are prevented from doing so by the same potential barrier at the metal surface which prevents their evaporation at room temperature. Thus the electrons are trapped in the wire by the potential barrier, and the force exerted upon the electrons is transmitted to the wire as a whole.

We infer from this experiment that a free electron (or ion) moving at right angles to a magnetic field would experience a force perpendicular to the field $\vec{H}$ and to the direction of the velocity $\vec{v}$. Since the electron is free to move, it will be accelerated in the direction of this force. What will be the shape of the path which the electron will follow if it moves at constant speed in a plane perpendicular to the magnetic field under the action of a force which is at all times perpendicular to $\vec{v}$?

The electron will experience a constant acceleration at right angles to its instantaneous velocity. This acceleration will, however, be constant only in magnitude. Its direction will change continually as the velocity changes direction because $\vec{a}$ has to remain perpendicular to $\vec{v}$. Such a situation is not new to us. We encountered it when we considered the motion of the moon about the earth. A body which is subject to a constant acceleration at right angles to its direction of motion executes uniform circular motion about a point at which the prolongations of the vectors of instantaneous (centripetal) acceleration intersect (Fig. 23.12). In other words, the acceleration vector $\vec{a}$, being perpendicular to $\vec{v}$, cannot change the magnitude of $\vec{v}$ but merely its direction. Owing to the constancy of $|a|$, the velocity changes its direction by a constant angular amount each second. The resulting path is a curve of constant curvature, namely, a circle.

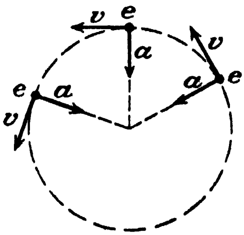


FIG. 23.12.

Figure 23.12 shows an electron e moving in a magnetic field which is perpendicular to the plane of the paper and is directed toward the reader.

The trajectory of an ion would have been likewise a circle. (The circular trajectory of charged ions in a magnetic field is made use of in the cyclotron. This device will be discussed in Chap. 39.)

Force Exerted upon Charged Particle Moving across Magnetic Field

We can now proceed to the quantitative question: How large is the centripetal force which compels the electron to move on the circular orbit? We can infer the answer to this question from the knowledge of the force F which a magnetic field exerts upon a wire through which a current is flowing. If the wire is at right angles to the magnetic field, F is given by the equation

$$F = \tfrac{1}{10}\mu HLi \tag{23.7}$$

According to the picture of the electron theory, F must be the sum of the forces exerted upon the n moving free electrons which are contained in the wire segment of length L. The problem of a constant electron flow through a wire can be treated in an analogous fashion to the problem of a steady flow of an incompressible fluid through a pipe. We have seen that the fluid volume which passes through the cross section of a pipe each second is equal to Av (where v is the fluid velocity and A is the cross-sectional area). If our fluid is a gas containing n molecules per cubic centimeter, (nAv) molecules will have passed the cross-sectional area in a second. Applied to our example of an electron gas flowing through a wire (Fig. 23.13), this means that nAv electrons are passing a cross section of our wire each second. Since each electron carries a charge of e coulombs, the charge passing through a cross section of our wire per second (*i.e.*, the current) must be

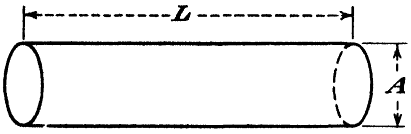


FIG. 23.13.

$$\boxed{i = Anve} \tag{23.8}$$

Substituting this value for i in equation (23.7) we get

$$F = \tfrac{1}{10}\mu HL(Anve) \tag{23.9}$$

But from Fig. 23.13 (LA) is the volume V of the length of the wire, n is the *electron density* or number of electrons per cubic centimeter; hence, $LAn = Vn = N$ is the number of electrons in the length L of the wire. Hence, we can write for equation (23.9)

$$F = \tfrac{1}{10}\mu HNev \tag{23.10}$$

From this we can compute the **force exerted upon an individual electron:**

$$f = \frac{F}{N} = \frac{1}{10}\mu Hev \tag{23.11}$$

If e is measured in abcoulombs, the factor $\frac{1}{10}$ is omitted:

$$\boxed{f = \mu Hev = Bev} \tag{23.11a}$$

Although we derived this equation by considering electrons, we actually made no assumptions which restrict the validity of this equation to the electronic charge e. Equation (23.11a) holds for particles of any charge e (measured in emu) moving with velocity v (in centimeters per second) at right angles to a field of intensity H (μ being equal to 1 in a vacuum). The force f is perpendicular to the velocity of the charged particle which moves in the magnetic field. It causes the centripetal acceleration which accounts for the circular motion of the particle. **The spatial relation between the directions of the current, force, and magnetic-field vector is indicated in Fig. 23.14. In the case of moving ions the velocity vector takes the place of the current vector $\vec{i}$.** (Compare this figure with Fig. 23.11.)

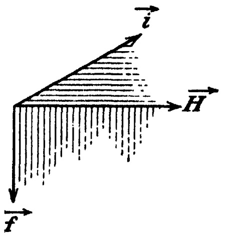


Fig. 23.14. This drawing is to be visualized in space (i is perpendicular to the plane of the paper and points away from the reader).

Equation (23.11a) allows us to define the meaning of the vector B as follows. We write $B = f/ev$. If we consider a unit charge moving with unit velocity, the magnitude of B becomes equal to the magnitude of f and we can state: *The magnitude of B is equal to the force exerted in the field upon a unit charge moving with unit velocity at right angles to the field lines* (*i.e.*, moving in such a direction that a maximum force is exerted upon it). The direction of B is perpendicular to the velocity of the charge and to the force exerted upon it. *Thus, whereas the magnitude of H is explored with a stationary magnetic pole, the magnitude of B is explored with a moving electric charge.*

The Circular Orbit in the Magnetic Field

Equation (23.11) enables us now to correlate the radius of the circular orbit of an ion (or an electron) moving in a magnetic field with its e/m. The centripetal force f given by equation (23.11) must be equal to $m(v^2/r)$, where m is the mass of the ion and v^2/r its centripetal acceleration. Thus we get in a vacuum

$$f = \frac{1}{10}Hev = m\frac{v^2}{r} \tag{23.12}$$

or
$$\frac{e}{m} = \frac{10v}{Hr} \text{ coulombs/gm} \tag{23.12a}$$

This expression would enable us to determine the "specific charge" e/m (charge per unit mass) of the given ion if we knew its speed v. And if, in addition, its charge e were known, the mass of the ion (or electron) could be determined in grams.

Ions of a known velocity can be obtained by accelerating them through a known potential difference V. The work done on an ion will be equal to eV, and its kinetic energy will be given by the equation

$$eV = \tfrac{1}{2}mv^2 \tag{23.13}$$

Hence
$$v^2 = \frac{2eV}{m} \tag{23.13a}$$

We can now eliminate the velocity v from equations (23.12a) and (23.13a). Squaring (23.12a) we get

$$\left(\frac{e}{m}\right)^2 = \frac{100}{H^2}\frac{v^2}{r^2} \tag{23.12b}$$

Substituting for v^2 the value from equation (23.13a) we get

$$\left(\frac{e}{m}\right)^2 = \frac{100}{H^2r^2}\frac{2eV}{m} = \frac{200}{r^2}\frac{V}{H^2}\left(\frac{e}{m}\right) \tag{23.14}$$

or
$$\underline{\frac{e}{m} = \frac{200}{r^2}\frac{V}{H^2}} \tag{23.14a}$$

The radius r of the ion's orbit, the accelerating voltage V, and the magnetic field H are all measurable quantities; hence, e/m can be determined experimentally.

Mass Spectrograph

One might think of using the following arrangement, set up in an evacuated space, for the determination of e/m of ions (Fig. 23.14a). S_0 is a source of positive ions some of which pass through the narrow opening in the screen S_1 after having been accelerated by the relatively small potential difference V_1 maintained by the battery B_1. The ions enter the interspace between S_1 and S_2 with velocities which are negligible as compared with the final velocity which they achieve due to acceleration by the potential difference V_2 between the plates S_1 and S_2. Thus ions which happen to pass through the opening in S_2 can be considered as having nearly equal velocities. When they emerge from the opening in S_2, they find themselves in a uniform magnetic field whose lines of force are perpendicular to the plane of the paper. They are bent by the magnetic field and proceed along a circular arc until they are stopped by the photographic plate *Ph*, which terminates their semicircular paths.† The distance of the points P_1 or P_2 from S_2 determines the diameter of the semicircle [and hence its radius R, which is needed for equation (23.14*a*)]. Thus, by determining the position of the mark P on the photographic plate and by measuring the field intensity H and the accelerating voltage V_2, we can establish the specific charge of our positive ions from equation (23.14*a*). The apparatus shown in Fig. 23.14*a* is called a *mass spectrograph.*

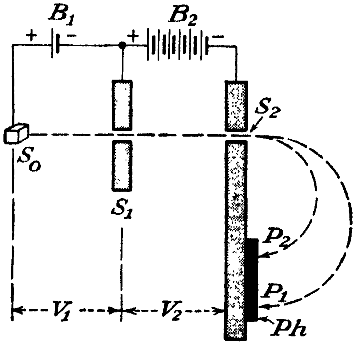


FIG. 23.14*a*. Scheme of a mass spectrograph. The magnetic field is perpendicular to the plane of the paper.

Isotopes

The study of the specific charge of ions of various elements led to a very important and surprising discovery.‡ In examining the positive rays of a given element, say boron, we should anticipate finding more than one dot P on the photographic plate, since we might expect the charge of some of the ions to correspond to more than one elementary charge. Thus we should look for a distinct black spot corresponding to boron ions whose charge is equal to 1e and quite far away for a much fainter spot corresponding to a charge of 2e. The actual experiment did reveal two spots with boron. But instead of being far apart as equation (23.14*a*) would predict for the charges e and $2e$, the spots were quite close together! The location of these spots on the photographic plate could, however, be well accounted for by assuming the existence of two kinds of boron atoms both carrying the same charge $+e$ but having slightly different atomic weights, namely, approximately 10 and 11. Tests with other elements showed similar results revealing frequently a great variety of atomic weights in the ion ray of a given element. For example, with tin 10 different spots are obtained corresponding to the atomic weights of 112, 114, 115, 116, 117, 118, 119, 120, 122, 124.

Thus mass spectrography has revealed that a chemical element is not a uniform substance containing atoms of identical mass but rather a mixture

---

† This design is a greatly simplified version of Dempster's early *mass spectrograph* model.

‡ Isotopes have been known to occur in a few cases from previous studies of radioactivity, but their general occurrence had not been demonstrated before the invention of the mass spectrograph.

of substances which all exhibit the same chemical properties but have different atomic masses. These substances are called **isotopes** of the given element. The per cent variation of mass from one isotope to another is very great in the case of the lightest elements, the greatest variation being encountered among the isotopes of hydrogen, which have the following atomic weights: 1 (hydrogen), 2 (deuterium), 3 (tritium).

The second important result obtained by mass spectrography was the discovery that **the atomic weight of all isotopes is very nearly an integral number** (*if the atomic weight of oxygen is assigned the atomic weight of precisely* 16). Elements whose atomic weight was known to differ appreciably from the nearest integer turned out to be mixtures of isotopes of nearly integral atomic weights. For instance, Cl, of atomic weight 35.46, turned out to be a mixture of isotopes of atomic weights 34.979 (75.4 per cent) and 36.978 (24.6 per cent).

The mass spectrograph can be used not only to measure e/m but also to separate isotopes of different mass from each other. This method has been used on a large scale in the separation of the U^{235} uranium isotope from its heavier isotope in the manufacture of the "atomic bomb."

Other methods of isotope separation are based on the difference between the rates of diffusion, of electrolytic migration (in the case of hydrogen), or of chemical reactions of the different isotopes.

The use of isotopes in biological research is of inestimable value. An element is usually a mixture of isotopes of different atomic weights, which occur together in a fixed proportion. For instance, ordinary hydrogen is a mixture of 99.98 per cent H^1 of atomic weight 1.008 and of 0.02 per cent H^2 (**deuterium** or **heavy hydrogen**) of atomic weight 2.015, or to cite another example, oxygen consists of 99.76 per cent of O^{16}, 0.04 per cent of O^{17}, and 0.20 per cent of O^{18}. If we now prepare, say, a fat, containing H^2 in a much higher concentration than it occurs in normal hydrogenous substances, we shall be able to determine the part of the body of the animal in which the food is deposited by establishing in what part of the body there is an abnormally high concentration of H^2. (The relative abundance of H^2 can be determined by means of a mass spectrograph.) In this example H^2 serves as a so-called **tracer.** Some fundamental biological discoveries have been made through the use of such tracer elements. We shall discuss these findings in more detail in connection with *radioactive* isotopic tracers.

The Specific Charge and the Mass of the Electron

The same principle as has been used to determine e/m of moving ions can also be used to determine the **specific charge of electrons.** Measurements performed with cathode rays yield the value

$$\boxed{\frac{e}{m} = 1.77 \times 10^8 \text{ coulombs/gm}} \tag{23.15}$$

Let us now compare the specific charge of the cathode-ray particles with the specific charge of, say, hydrogen ions. We know that 1.008 gm of hydrogen (1 gm equivalent) are transported along with 96,500 coulombs of charge. Hence, the **specific charge of hydrogen ions must be**

$$\boxed{\left(\frac{e}{m}\right)_H = \frac{96{,}500}{1.008} = 9.58 \times 10^4 \text{ coulombs/gm}}$$

For all other ions e/m will be smaller, since hydrogen ions have the smallest mass of all the ions.

We see that the specific charge of the electrons is about 2,000 times (more precisely 1,840 times) as high as that of hydrogen ions. Does this mean that cathode rays consist of particles of a mass close to the mass of hydrogen atoms carrying an enormous charge of 1,840 elementary units, or do they consist of particles carrying one elementary unit having a mass of 1/1,840 of the mass of a hydrogen atom? (These are, of course, not the only two possible assumptions about the values of e and m, there being an infinite number of possibilities consistent with our value for e/m.) The decision can be made only by an independent measurement of e or of m. As we know, the value of a tiny charge e can be determined by Millikan's method. Such measurements yielded the aforementioned value of $e = 1.60 \times 10^{-19}$ coulomb for the charge of the cathode-ray particles. This means that their charge is the same as the charge of hydrogen ions but their mass is 1,840 times smaller.

Substituting Millikan's value of e in $(e/m)_H = 1.60 \times 10^{-19}/m_H = 9.58 \times 10^4$ coulombs/gm, we get

$$\boxed{m_H = 1.66 \times 10^{-24} \text{ gm}}$$

for the **mass of a hydrogen atom,** which agrees with the value found through the kinetic theory of gases.

If we substitute Millikan's value of e in equation (23.15), which applies to (slow) cathode-ray particles, we find

$$\boxed{m_0 = 8.98 \times 10^{-28} \text{ gm}}$$

for the **mass of an electron.** Thus, cathode rays consist of *particles carrying a negative elementary electric charge and having a mass* 1,840 *times as small as the mass of the lightest known atom.* These results are obtained with all cathode rays, no matter how they are produced and from what substance they emanate.[†]

We have now established a second property of electrons. We know their mass in addition to their charge. These data were obtained by the study of free electrons in a vacuum, and the question now arises whether or not we could demonstrate the presence of charged particles having the above-mentioned charge and mass in the interior of metals to substantiate our hypothesis about the electronic mechanism of metallic conduction. The fact that electrons can be "evaporated" from a metal surface does not prove that negative electricity is present in the form of electrons inside the metal. When we observe water dripping out of a leaky container, escaping in the form of round drops of a definite size, we do not conclude that water inside the container is composed of spherical droplets of the same size as those which escape. Through an ingenious scheme, Tolman actually succeeded in determining the specific charge of the carriers of electric currents in the interior of metals.

Tolman's Experiment. Imagine a metal cylinder accelerated in the direction of its axis with the constant westward acceleration $\vec{a}$. To an observer accelerated along with the cylinder all particles inside the cylinder will appear to have a weight $-(\overrightarrow{ma})$, which is a force directed eastward. Now if there are electrons in our metal cylinder, each of them will appear to an observer who shares the acceleration of the conductor to be acted on by a force $\vec{F} = -(\overrightarrow{ma})$. If the electrons are free to move, they will be accelerated *eastward* relative to the conductor. But as the eastward end of the cylinder gains electrons while the westward end becomes depleted of electrons, an electric field is set up within the cylinder, which counteracts the migration of the electrons toward the eastward end of the rod. The migration will stop when the field E has become so strong that the westward electrostatic force exerted upon an electron has become equal to the eastward d'Alembert force:

$$\text{Electrostatic force} \rightarrow Ee = -ma \leftarrow \text{d'Alembert force} \qquad (23.16)$$

[†] e/m changes with the speed of the cathode rays. This effect becomes important when the speed of the electrons approaches the speed of light but can be disregarded at slow speeds. This phenomenon will be explained in Chap. 40.

If the length of the rod is L and the potential difference across the rod is V, we can write for the field strength inside the metal

$$E = \frac{V}{L} \tag{23.17}$$

Substituting equation (23.17) in (23.16) we obtain

$$\frac{V}{L}e = -ma \tag{23.18}$$

or

$$\underline{e/m = -\frac{aL}{V}}$$

We see that, if we measure the acceleration a, the length of the rod L, and the potential difference developed across its ends, *we can get* e/m *for the carriers of the electric charge inside the metal.*

Tolman performed such an experiment and found a value of e/m which agreed within the limits of experimental error with e/m for slow cathode rays. We can accept this result as a strong supporting evidence for the assumption that there are freely movable electrons in metals.

Magnetism as an Electric Phenomenon

We began this chapter with the discovery of the relationship between electricity and magnetism. The study of the magnetic manifestations of electric currents led us eventually to the determination of the specific charge and of the mass of the electron, and the knowledge of those data made it possible to provide experimental evidence for the existence of free electrons in the interior of metals. All electric and electromagnetic phenomena which we observed with metallic conductors can now be interpreted as being due to the motion of electrons. Electrons can be made to drift by maintaining an electric field inside the metal by means of a battery. This orderly electronic motion is associated with a magnetic field. In order to explain the thermionic effect ("evaporation" of electrons from hot metals), we assumed the electrons to be in a persistent state of random motion. It is, however, easy to see that this random motion cannot give rise to a magnetic field, since we could find for every electron A another electron B in its neighborhood moving in the opposite direction so as to produce a magnetic field canceling the magnetic field produced by electron A. Is it, however, not possible to imagine some kind of an orderly, spontaneous electronic motion which could persist in the interior of a metal, giving rise to a net magnetic field? Would it, perhaps, be possible to interpret the magnetic properties of a permanent magnet by such electronic motion? If such a possibility exists, magnetism will not have to be considered any more as an isolated physical phenomenon but could, rather, be interpreted in all cases as a manifestation of motion of electrical charges. The following discussion will show that such an interpretation of magnetism is indeed possible.

Let us begin with a comparison between the magnetic field of a flat coil and that of a flat magnet. The lines of force, which can be established experimentally by means of iron filings, are indicated in Fig. 23.15*a* (for a loop), *b* (for a coil), and *c* (for a magnet). The magnetic field of the loop and the coil resembles the field of a bar magnet. In fact, if a loop or a coil which is traversed by a current is suspended like a compass needle, it orients itself so that the central axis points from north to south; *i.e.*, the plane of the loop sets itself at right angles to the field. The great similarity between the magnetic fields in these cases suggests that a coil traversed by a current may be considered as being equivalent to a magnet or, conversely, that the field of a magnet may be considered as being due to circulating currents which somehow exist in the magnetized iron bar. Although the last statement seems to make hardly any sense at first glance, we shall see presently that it constitutes a fruitful idea which offers a plausible explanation for the puzzling fact that it is not possible to isolate the north pole of a magnet from its south pole.

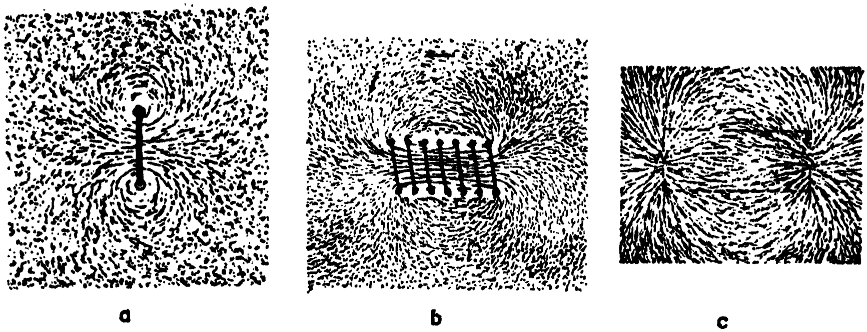


FIG. 23.15. *a.* The magnetic field of a loop. *b.* The magnetic field of a coil. *c.* The magnetic field of a bar magnet. *(From Pohl, "Einführung in die Elektrizitätslehre," Springer.)*

Our inability to separate magnetic poles of opposite sign as we have been able to separate opposite electric charges constitutes the most pronounced dissimilarity between electric and magnetic phenomena. We have sought to explain this difference by assuming magnetization to be due to orientation of permanent magnetic dipoles whose poles are bound so firmly to each other that, when we break a magnet, they don't break apart. But this should still leave the hope that by appropriate means a breaking up of the elementary magnetic dipoles could be accomplished. Failure to achieve this aim led Ampère to the belief that magnetic poles (*i.e.*, particles with a "magnetic charge") do not exist.† Ampère suggested that the "molecular magnets," or permanent dipoles (the orientation of which gives rise to the magnetization of a body), are actually tiny **molecular currents.** In other words, we should imagine (without taking this picture literally) tiny conducting loops to exist within the atoms in which electronic currents circulate forever.‡ Since a current loop has no material north and south poles, this hypothesis explains why it is impossible to separate and isolate the opposite magnetic poles by breaking up individual molecules and atoms.

Effect of Iron on the Magnetic Field

This hypothesis also offers an explanation of the increase of the magnetic field strength which takes place near a coil through which a current is flowing when we insert an iron bar into the coil. Let us attempt to explain the following experiment in terms of Ampère's hypothesis of molecular currents: A magnetic test pole N is placed at the point P at some distance from a coil C through which a constant current is flowing (Fig. 23.16*a*). The force exerted upon N in the magnetic field of the coil is determined. Experiments show that this force can be augmented either by increasing the current or by increasing the number of turns of the coil. There is, however, another way of increasing this force, which does not involve a change in current or in the number of turns. When an iron bar B (Fig. 23.16*b*) is inserted into the coil, the force upon N is

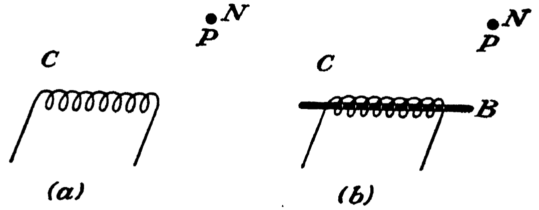


FIG. 23.16.

---

† Recently, Dirac, who predicted the existence of positive electrons, has suggested a hypothesis postulating the existence of magnetic poles. The force of attraction between two opposite Dirac poles is assumed to be so great that it would be very difficult to separate them.

‡ According to modern theories the revolution of electrons plays a role in the magnetism of nonferrous materials (paramagnetism) along with the *rotation* of the electrons. The magnetism of iron (ferromagnetism) is ascribed primarily to the electron spin (rotation).

greatly increased. It may become hundreds of times as large as it was without iron, as if many new turns have been added to the coil.

This phenomenon can be interpreted as follows, according to Ampère's hypothesis: Assume that the magnetizing field H ($\overrightarrow{H}$ being perpendicular to the plane of the paper pointing away from the reader) orients the molecular current loops as shown in cross section in Fig. 23.17. (For greater simplicity of explanation the loops are drawn rectangular in shape, although, as we shall see in Chap. 36, the circular shape is the reasonable one to assume.) We see that the currents in the adjacent sides of any two neighboring loops are opposed to each other, and if we assume them to be equal, we see that the magnetic fields of the currents in the interior of the iron bar (the long side of which is parallel to $\overrightarrow{H}$, *i.e.*, perpendicular to the plane of the paper) cancel out. The only currents whose magnetic fields remain uncanceled are those flowing in the portions of the loops which lie along the boundary of the magnetized bar (Fig. 23.17). The effect of these uncanceled *external* currents is magnetically the same as that of a current circulating around the surface of the bar as shown in the diagram by the large square of arrows. Thus the iron bar inserted into a solenoid through which a current flows behaves as if electric currents were circulating around its surface, in other words, as if invisible turns of wire were wound on its surface carrying a current in the same sense as the turns of the magnetizing solenoid. Thus our solenoid with an iron core inserted through it behaves like a solenoid with an increased number of turns, and hence, its original magnetic flux is increased by the flux lines contributed by the oriented molecular currents (or, which is the same, by the fictitious surface currents referred to above). The same hypothesis can be used to explain qualitatively the magnetic behavior of paramagnetic substances.

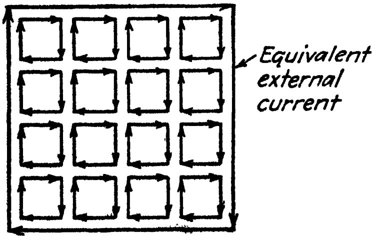
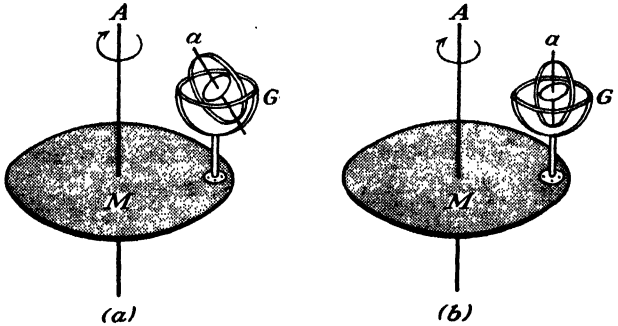


FIG. 23.17.

Despite its attractive features, the hypothesis of molecular currents seems to be hard to accept. We know that iron has a rather high electrical resistivity, and we know that iron gets hot when we pass a current through it. And yet a magnetized iron bar does not get hot despite the circulating currents we assume to exist in it. We have to suppose that the mechanism of flow of the "molecular currents" is different from the passage through the metal of an electron current which is maintained by a battery connected across the conductor.

Barnett Effect as Experimental Support for Ampère's Hypothesis of Molecular Currents

The simplest assumption which explains how molecular currents could exist without producing heat is to visualize electrons revolving in empty space inside an atom. Evidence for the support of this idea was provided by Barnett through an ingenious experiment. Barnett reasoned as follows: Since an electron has mass, a revolving electron will possess an angular momentum. It should behave similarly to a disk rotating about a central axis perpendicular to its plane. Such a system, endowed with angular momentum, constitutes a simple example of a so-called *gyroscope* (or top). A spinning gyroscope possesses the following interesting property: When placed on a rotating merry-go-round *M* in such a way that the axis *a* of the gyroscope *G* is not parallel to the axis of rotation *A* of the merry-go-round (Fig. 23.18*a*), the axis of the gyroscope tends to set itself parallel to the rotational axis of the merry-go-round (Fig. 23.18*b*). This phenomenon underlies the principle of the gyrocompass, which consists essentially of a fast-spinning gyroscope whose axis of rotation orients itself in the plane of the geographic meridian owing to its tendency to orient itself as nearly as possible parallel to the axis of rotation of the earth. Barnett conceived of the billions of iron atoms containing circulating (revolving as well as spinning) electrons as tiny gyroscopes. He expected that by spinning a cylindrical iron bar about the cylinder axis the axes of the tiny "atomic gyrocompasses" would all orient themselves parallel to the axis of rotation. This, however, would be equivalent to orientation of the molecular magnetic dipoles. Thus, *our picture of circulating electrons predicts the possibility of magnetizing an iron bar by rotating it.* Barnett's experiment proved successful and thus confirmed the belief of physicists in the existence of circulating electrons in atoms.



(a) (b)

FIG. 23.18.

APPENDIX I

Electrical Units (Conversion Table)

	Electrostatic units	Electromagnetic units	Practical units
Charge.......	1 esu $= \frac{1}{3 \times 10^{10}}$ abcoulomb $= \frac{1}{3 \times 10^{9}}$ coulomb	1 emu = 1 abcoulomb $= 3 \times 10^{10}$ esu of charge = 10 coulombs	1 coulomb $= \frac{1}{10}$ abcoulomb $= 3 \times 10^{9}$ esu of charge
Current......	1 esu $= \frac{1}{3 \times 10^{10}}$ abamp $= \frac{1}{3 \times 10^{9}}$ amp	1 emu = 1 abamp $= 3 \times 10^{10}$ esu of current = 10 amp	1 amp $= \frac{1}{10}$ abamp $= 3 \times 10^{9}$ esu of current
Potential difference.....	1 esu of p.d. $= 3 \times 10^{10}$ emu of p.d. = 300 volts	1 emu of p.d. $= \frac{1}{3 \times 10^{10}}$ esu of p.d. $= 10^{-8}$ volt	1 volt $= 10^{8}$ emu $= \frac{1}{300}$ esu of p.d.

APPENDIX II

Electrical Measuring Instruments

The Moving-coil Galvanometer

The force exerted upon a current-carrying conductor in a magnetic field is utilized in the design of the moving-coil **galvanometer,** a sensitive device for measurement of currents. Figure 23.19 illustrates the principle of this instrument. The current which is to be detected or measured flows through a flat coil consisting of many loops, one of which is shown in the figure as a frame through which a counterclockwise current is flowing. The coil is suspended from a fine wire S (which forms one of the leads through which the current is supplied to the coil) in a magnetic field (the second lead wire is not shown). No force is acting on the horizontal wires U and D through which the current flows parallel to the magnetic lines of force. The vertical wires R and L experience a force [given by equation (23.7)] at right angles to the plane of the paper in Fig. 23.19. The diagram shows how the two forces F and $-F$ give rise to a torque about the axis of suspension. In order to avoid diminution of the torque, which occurs when the coil is rotated so that its plane ceases to be parallel to the magnetic lines of force, an iron cylinder is

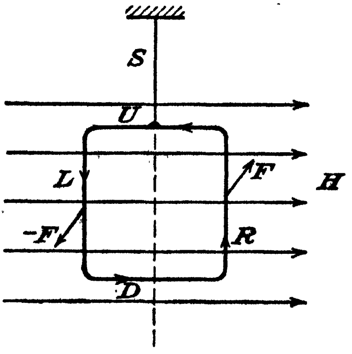


FIG. 23.19. Scheme of a moving-coil galvanometer.

introduced into the coil so as to make the magnetic field radial. Then as the coil rotates, its plane remains parallel to the magnetic lines of force which pass through the vertical wires. This method of achieving a nearly radial magnetic field is shown in top view in Fig. 23.20. C is the cylinder placed inside the coil, U is the top section of the coil frame, and N and S are the field-producing poles.

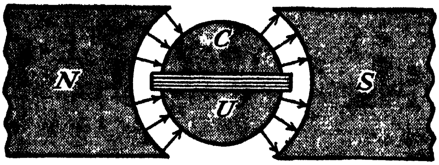


FIG. 23.20. Magnetic field of a moving-coil galvanometer. N and S, magnetic pole pieces; C, iron cylinder; U, moving coil in top view.

Owing to the electromagnetic torque the suspension is twisted through an angle which can be read on a scale by means of a pointer attached to the moving coil (or better, by means of a light beam reflected from a mirror attached to the coil). The angular excursion of the coil is proportional to the current which can thus be measured. Such instruments can be made sensitive enough to detect currents in the order of magnitude of 10^{-10} amp.

The Voltmeter

For practical, more rugged instruments, the fine wire suspensions are replaced by a mounting in jewel bearings. The restoring torque is then provided by a special spring (Fig. 23.21).

In measuring a potential difference between two points one wishes to draw as little current as possible in order not to change appreciably the magnitude and distribution of the current in the circuit by the introduction of the measuring instrument. Hence, a voltmeter is composed of a sensitive galvanometer with a high resistance placed in series (see Prob. 18, Chap. 22). Thus only a small current passes through the instrument even though a high voltage may exist across its terminals (Fig. 23.22a).

The Ammeter

An ammeter is an instrument for the measurement of current which is inserted into the branch of the circuit where the current is to be measured. In order to avoid a great change of the branch resistance by the introduction of the instrument, it is desirable to make its resistance as low as possible. This is accomplished by the use of a sensitive galvanometer G, which is "shunted" by a low resistance S (Fig. 23.22b).

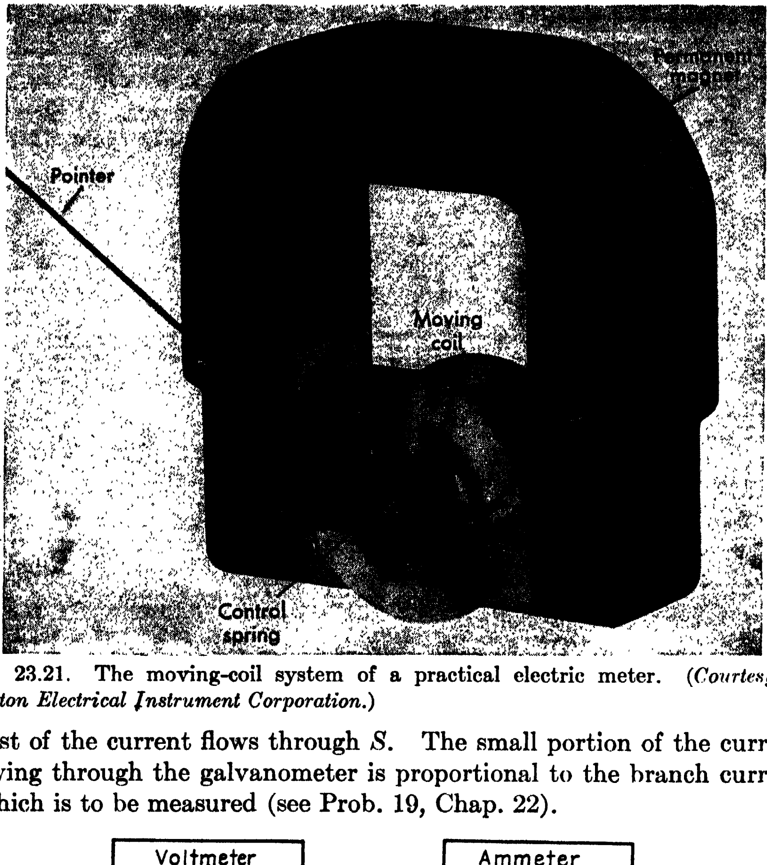


FIG. 23.21. The moving-coil system of a practical electric meter. (*Courtesy of Weston Electrical Instrument Corporation.*)

Most of the current flows through S. The small portion of the current flowing through the galvanometer is proportional to the branch current i which is to be measured (see Prob. 19, Chap. 22).

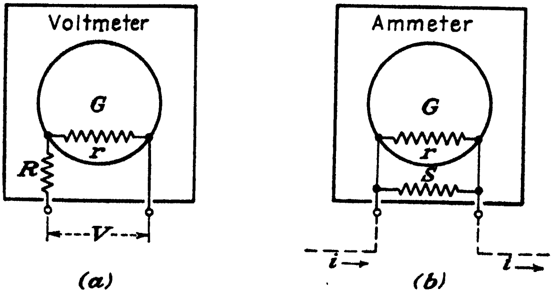


FIG. 23.22*a*. Scheme of a voltmeter. G, galvanometer; R, series resistor; r, resistance of the galvanometer coil. *b*. Scheme of an ammeter. G, galvanometer; S, shunt; r, resistance of the galvanometer coil.

The Cathode-ray Oscilloscope

Biologists use a smoked-drum kymograph (Fig. 23.23) to record variable quantities, such as a fluctuating blood pressure. The stylus

S of **Fig. 23.23** is attached to a thin rod, which is supported by a small float floating on top of the mercury of the manometer M. As the drum turns, a record of the blood pressure as a function of time is made.

In many instances of physical or biological applications we are interested in recording or visualizing the variation of an electrical quantity, for instance, of voltage as a function of time. For example, the recording of an "action potential" of an active nerve fiber requires the recording of a transient variation of a potential difference which may take place within 10^{-4} sec. This is made possible by the **cathode-ray oscilloscope** shown in Fig. 23.24. Electrons originate in the hot filament C, pass through the small opening in the diaphragm A, and proceed as a fine beam B toward the fluorescent screen S, where they give rise to a luminous spot E at the point where they strike the screen. A potential difference between the plates P_1' and P_1'' creates a horizontal electric field which causes a horizontal displacement of the spot E. This displacement is due to the horizontal acceleration which the electrons experience as they pass between these plates. Similarly, a potential difference between P_2' and P_2'' causes a vertical displacement of the spot E.

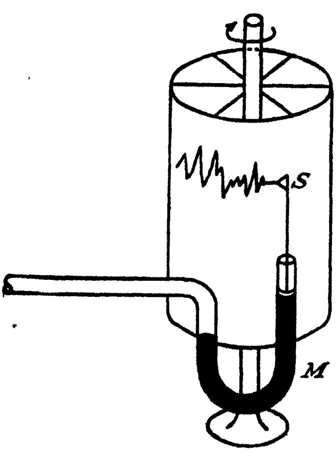


FIG. 23.23. A smoked-drum kymograph.

This device can be used as a voltmeter as follows: If we connect a potential difference across the "vertical deflection plates" P_2' and P_2'', the electron beam B will be deviated in the vertical direction and the position of the fluorescent spot E changes so that the device can be calibrated to read the applied potential difference from the observed displacement of the point E. The vertical deviation of the luminous spot E due to an applied voltage is analogous to the vertical deviation of the stylus S in Fig. 23.23, which measures blood pressure. Now if we could move the screen S relative to the electron beam horizontally, we should have a complete analogy to the kymograph. The desired effect can be accomplished easily by leaving the screen at rest but "sweeping" the electron beam horizontally by a variable potential difference applied

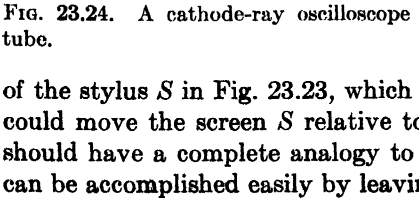


FIG. 23.24. A cathode-ray oscilloscope tube.

to the horizontal deflection plates P_1' and P_1''. If the horizontal potential difference is varied so that it increases proportionally with time, the spot produced by the electron beam on the screen S will move with a uniform velocity horizontally along the screen in the absence of a vertical deflection.† If, now, the variation of the potential difference across a nerve fiber is applied to the vertical deflection plates,‡ the spot will move vertically, its vertical position indicating the instantaneous applied voltage (the "action potential" of the nerve fiber) while at the same time moving uniformly to the right and thus depicting a curve which represents the varying "action potential" as a function of time (Fig. 21.10).

The chief advantage of this instrument lies in the fact that, owing to the small inertia of the electrons, it can be used to record faithfully very rapidly varying electrical processes.

This tube also forms a basic part of a **television receiver.** Imagine the electrons to be swept horizontally across the screen in $\frac{1}{20}$ sec, while at the same time the vertical deflection plates sweep the spot, say, 10^4 times up and down so that the luminous spot sweeps over the entire screen in $\frac{1}{20}$ sec and produces the illusion of uniform illumination of the entire screen. If a skillful artist were capable of controlling the luminous intensity of the spot with unlimited speed during this $\frac{1}{20}$ sec, he could vary it so as to paint a black-and-white picture on the screen by dimming and brightening the spot E as it passes various points on the screen. This kind of control can actually be accomplished by a television transmitter, which plays the role of the artist in our example.

APPENDIX III

The D-c Motor

The principle of a motor is quite similar to the principle of the galvanometer illustrated in Fig. 23.19. The motor, too, possesses a coil through which a current is fed. But the angle through which the coil of the motor can rotate is not limited. In Fig. 23.19 the coil turns counterclockwise looking from the top at the wire U. When its plane reaches a position perpendicular to the plane of the paper, the torque is zero, since then the two forces F and $-F$ both lie in the plane of the coil and pull in opposite directions along the same line. But the coil has acquired rotational kinetic energy and "overshoots." It now should be turned back in the opposite sense, since the forces F and $-F$ did not change their direction. Thus instead of continuing to rotate counterclockwise the coil would oscillate about a plane perpendicular to the plane of the paper. This is avoided by the use of a *commutator* (Fig. 23.25). The terminals of the coil are attached to two isolated metallic semicircles S_1 and S_2. The current is supplied to the coil through the brushes B_1 and B_2, which slide along the semicircular terminals as the coil is rotating about the axis AA. When the coil reaches the position perpendicular to the one shown in the diagram, it "overshoots," continuing to rotate in the same sense The half rings S_1 and S_2 now change their connection. Instead of making contact with B_1, S_1 now touches B_2 and S_2 touches B_1. As a result, the current in the coil is now reversed, and hence, instead of suffering an angular retardation after overshooting, it continues to experience a torque moving it in the same direction of rotation.

---

† The spot is "swept" with uniform speed across the screen and then returned quickly to zero. This sweeping motion is repeated periodically.

‡ In practice this variation of potential difference must be amplified electronically before it is large enough to cause a visible deflection of the cathode-ray beam.

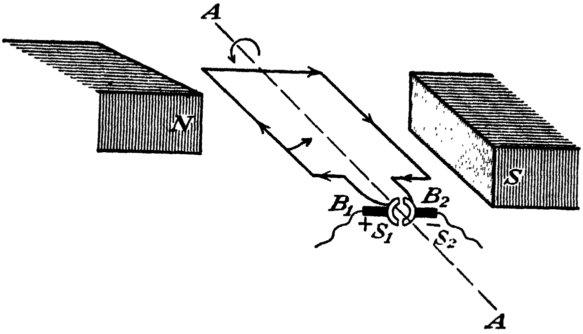


FIG. 23.25. Scheme of a d-c motor.

QUESTIONS AND PROBLEMS

1. How do the equipotential lines look in cases where the lines of force are circular?

2. What difficulty do you encounter when you attempt to define a magnetic potential at a point near a wire traversed by a current?

3. What is the value of Biot's law? Nobody can obtain an isolated current element ΔL traversed by a steady current.

4. How did Biot infer his law? Did he find it directly by experimentation, or did he guess it and verify its correctness by agreement of its predictions with experience?

5. How large is the force which a magnetic north pole of 10 cgs units placed at the center of a circular coil of five turns 10 cm in radius, exerts upon the coil, which carries a current of 20 amp?

★6. Can you propose a way, different from the method mentioned in the text, for the definition of an emu of current? (Your method should not make use of magnetic poles.)

7. How much work is done in carrying a N pole of 20 cgs units three times completely around in a circle of 5 cm radius near a long straight wire carrying a current of 30 amp? The plate of the circular path is perpendicular to the wire, and the wire does *not* pass through the area of the circle.

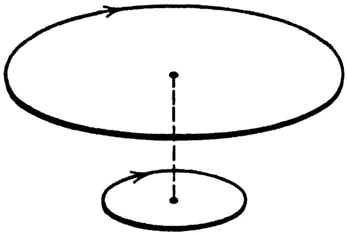
★**8.** Can you derive the value of the field intensity in the interior of a toroidal solenoid (*i.e.*, a coil wound on a doughnut)?

9. How can you influence the magnetic field produced at some distance from a coil without changing the current passing through it?

10. Can you find the force in magnitude and direction exerted by the earth's magnetic field upon a current of 1,000 amp in a wire 200 cm long held parallel to the lines of force of the earth's magnetic field?

11. Which way will a wire carrying an east-west current be deviated in a north-south magnetic field?

12. Will two parallel circular loops whose centers lie along a line perpendicular to the plane of the loops ($R_1 = 20$ cm; $R_2 = 10$ cm) carrying clockwise currents i_1 and i_2 (as seen from the top) (*a*) attract each other? (*b*) repel each other? (*c*) exert a torque upon each other? (*d*) tend to rotate about the axis joining their centers?



13. Where do electric lines of force originate, and where do they end?

14. Where do the magnetic flux lines of a bar magnet originate and end?

15. Where do the magnetic flux lines of a coil originate, and where do they end?

16. Do you consider an electromagnet preferable to a permanent magnet for the purposes of extracting magnetic foreign bodies from eyes? Why or why not?

★**17.** *a.* If a soft iron bar attracts a bar magnet, does it follow from this that a soft iron bar will attract another soft iron bar?

b. If a wire traversed by an electric current exerts a force on a magnetic pole, does it follow from this that it will also exert a force upon another wire traversed by a current?

18. The positive and negative ions in a pipe filled with an electrolyte through which a current is passed move in opposite directions. Will such a pipe experience a force when placed in a magnetic field whose field lines are perpendicular to the current?

19. Why does a magnetic field not force the electrons out of a wire in which a current is flowing at right angles to the magnetic field? What hypothesis would you suggest to explain this observation?

20. What is the trajectory of a charged particle that moves at constant speed at right angles to the direction of a homogeneous magnetic field? What is the reason for this shape of the trajectory?

21. How much work is done in carrying a magnetic pole of 20 cgs units three times around a wire carrying a current of 25 amp?

22. The fiber of an electrocardiograph galvanometer is 5 cm long, and the strength of the magnetic field, which is perpendicular to the fiber, is 20,000 oersteds. If the resistance of the metallized quartz fiber is 5,000 ohms, how large a force will be exerted upon it when a p.d. of 10^{-3} volt is applied across the fiber?

23. Why is a force exerted upon a charge when it moves through a magnetic field in a direction perpendicular to $\vec{H}$, whereas no force acts upon it when it is at rest with respect to the field lines?

★**24.** Does an electron moving in a plane perpendicular to $\vec{H}$ in a uniform magnetic field in a circular orbit influence the intensity of the magnetic field inside its orbit?

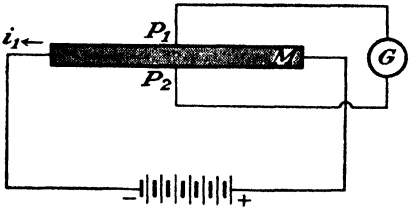
25. How do you explain the fact that the flux inside a coil changes when an iron core is inserted?

26. How do you explain the fact that we have not succeeded in isolating magnetic north poles from south poles?

27. What objections can you cite against Ampère's hypothesis of molecular currents? What evidence can you cite in its favor?

28. How can isotopes be used in biological research?

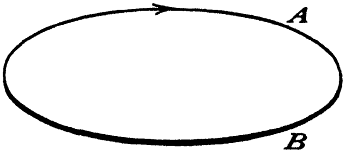
29. A current is passed through a conducting metal strip M. The points P_1 and P_2 are equipotential points. What will you expect to observe in the illustrated setup when a strong magnetic field is established at right angles to the plane of the paper?



30. An oxygen (O^{16}) ion carrying two elementary positive charges is accelerated from a negligible initial speed to a high final velocity by "falling" through a p.d. of 100,000 volts in an electric field. The rapidly moving ion then enters a magnetic field of 10,000 oersteds. How large a force is exerted upon this ion in the magnetic field if it moves at right angles to the lines of force?

31. Which of the following statements is correct? The experiment on deviation of cathode rays in a magnetic field proves that (*a*) they consist of electrons, (*b*) they consist of moving negative electricity, (*c*) electrons are discrete particles, (*d*) the charge of an electron is 1.59×10^{-19} coulomb.

32. If this current loop were replaced by a flat bar magnet, where would be the north pole of the magnet?



33. The dimensions of a galvanometer coil, such as shown in Fig. 23.19, are as follows: $L = R = 4$ cm; $U = D = 2$ cm. There are 100 turns of wire wound on the frame, and the magnetic field strength is 5,000 oersteds. Compute the torque exerted on the coil when $i = 10^{-8}$ amp.

34. Two parallel wires, each 6 m long, are spaced 2 cm apart. Compute the magnetic field intensity at the center between the wires when (*a*) both carry a current of 10 amp in the same direction, (*b*) when both carry a current of 10 amp in opposite directions.

Also compute the force (attraction or repulsion?) between the wires in the cases (*a*) and (*b*).

35. A hydrogen ion is moving with a speed of 10^8 cm/sec in a magnetic field of 10,000 oersteds at right angles to the lines of force. Find the radius of its orbit.

36. How fast would an electron have to move in a field of 10,000 oersteds at right angles to the magnetic field in order to describe a circle of 1 cm radius?

★37. Design a flat coil 10 cm in diameter which could be connected to a 110-volt battery and develop a field of 100 oersteds at its center without undue heating. (Use your judgment to decide what is admissible heating.)

BIBLIOGRAPHY

Urey, H. C., Separation and Use of Stable Isotopes, *J. Applied Phys.*, Vol. 12, p. 270, 1941.

Schoenheimer, R., and S. Ratner, Metabolism of Proteins and Amino Acids, *Ann. Rev. Biochem.*, p. 197, 1941. (Describes the use of stable isotopes as tracers.)

CHAPTER 24

INTERACTION BETWEEN MAGNETIC FIELDS AND MOVING UNCHARGED BODIES

a. Under what conditions is an emf induced in a conductor?

b. Is electromagnetic induction necessarily accompanied by the flow of induced currents?

c. Is an emf induced only in metallic conductors?

d. What do we mean by the statement: "The conductor moves so as to cut magnetic lines of force"?

e. How would you dispose a secondary coil with respect to a primary coil in order to make the emf induced in the secondary as high as possible?

f. Is it possible to induce a constant emf in a coil? How or why not?

g. On what factors does the magnitude of the emf induced in a secondary coil depend?

h. How can you predict the direction of the current induced in a loop or a coil?

i. How is it possible to link by a common description the induction of an emf in a closed circuit by a changing magnetic flux with the induction of an emf due to relative motion between a magnetic field and a conductor?

j. Where does the energy of the induced current come from?

k. What is self-induction?

l. What is inductance?

m. How can you predict the direction of the self-induced current?

n. Where does the energy of the self-induced current come from?

o. What is a "Lorentz field"?

p. How do you picture the mechanism of electromagnetic induction according to the electron theory of metallic conduction?

q. How do you explain electromagnetic induction in a flowing salt solution?

r. What is the principle of the betatron?

We have seen in Chap. 23 that a moving electric charge is associated with circular magnetic lines of force (Fig. 23.7*a*). We become aware of this magnetic field when the charge passes a stationary magnetic pole (Fig. 23.7*b*). The pole then experiences an impulsive force in the direction of the local magnetic lines of force. Conversely, should we get an electric field around a moving magnetic pole such that, if the moving charge in Fig. 23.7*a* were replaced by a magnetic pole, the circular lines of force would be *electric* lines surrounding the moving magnetic pole? How could we detect such a field if it did exist? This should be easy. If the magnet pole N of Fig. 23.7*b* were replaced by an electric charge, this charge should experience a force and should be accelerated in the assumed electric field of the moving magnetic pole. Now there are freely movable electrons in the interior of metals, and thus we should expect a current to be generated in a metallic conductor which is passed by a fast-moving magnetic pole if the moving pole is indeed associated with an electric field.

Faraday's Discovery

Such a process of *induction of electrical currents by moving magnets* was actually discovered by Faraday (**electromagnetic induction**). Faraday observed a momentary deflection of a galvanometer G (indicating a transient current) when one of the poles of a bar magnet was moved in or out of a coil (Fig. 24.1).

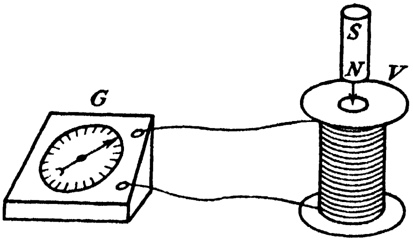


FIG. 24.1. Demonstration of electromagnetic induction of an emf in a coil.

1. Coil Stationary, Pole Moving. Evidently *an emf had been induced in the coil.* This emf persisted as long as the magnet was moving. This was evidenced by the fact that every motion of the magnet relative to the coil was accompanied by an electric current. The magnitude of the galvanometer deflection increased as the velocity of the magnet increased, and the sense of the galvanometer deflection was reversed when, instead of a north pole, a south pole was inserted. For the motion of a given pole the sign of the deflection was reversed by reversing the direction of motion. In other words, if a clockwise deflection resulted from the insertion of the north pole, a counterclockwise deflection was observed while the north pole was being removed.

These observations confirm our hypothesis that an electric field is associated with a moving magnetic pole. But Faraday observed more than we have predicted on the basis of this assumption.

2. Coil Moving, Pole Stationary. Faraday observed that precisely the same deflection was observed when the coil was moved toward the stationary magnet as was recorded when the magnet was moved toward the coil with the same speed. Quantitative observations showed that the magnitude of the galvanometer deflection (and, hence, the magnitude of the induced emf) depends only on the *relative velocity between the coil and the magnet.*

3. No Change of Distance between Coil and Magnet. But the induction of an emf in a coil or a loop does not necessarily require a relative change of distance between coil and magnet. For instance, we can rotate the flat coil of Fig. 24.2 from position 1 to position 2 and record an induced emf during this change in orientation. Or we can change the area of the coil or loop, for instance, by deforming an originally

0-shaped loop into an 8-shaped loop. This diminution of loop area (Fig. 24.3) also gives rise to an induced emf.

4. Both Coil and Magnet Motionless. Finally, an emf may be induced in a coil by a magnet without either of the two bodies being moved. Figure 24.4 shows a coil S facing an electromagnet P, *i.e.*, a coil through which a current is flowing. When the switch Sw is closed or opened, the current through P is initiated or discontinued. In both cases the galvanometer G gives a transient deflection. If the sense of

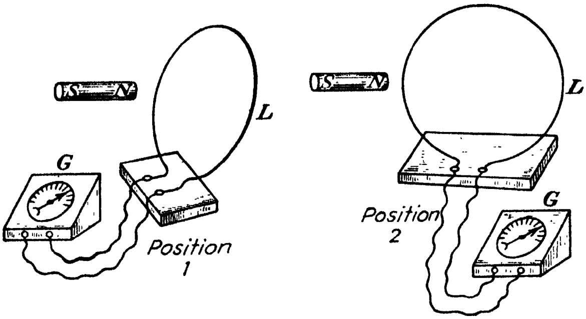


FIG. 24.2. An emf is induced in the loop L by rotating it with respect to the magnet from position 1 to position 2.

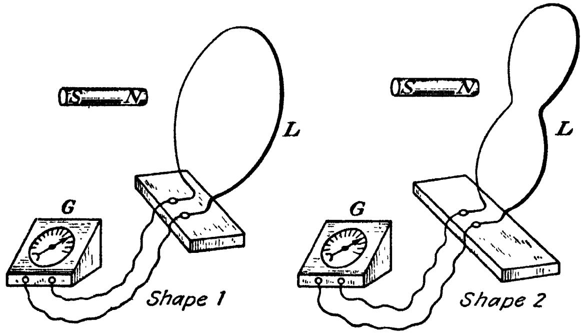


FIG. 24.3. An emf is induced in the loop L as its area is diminished by deformation.

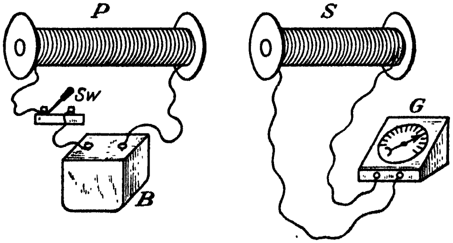


FIG. 24.4. Induction of an emf in the secondary coil S by starting or stopping the current flow through the primary coil P.

the deflection is clockwise when the switch is closed, a counterclockwise deflection results when the switch is opened. Electromagnetic induction can also be observed with a gradual increase and gradual decrease in current intensity instead of the sudden changes obtained by closing and opening the switch. Figure 24.5 shows how such gradual changes in i can be produced by a variable resistance R in series with the coil P. As R is diminished, i increases and so does the magnetic field generated by the coil P. At the same time the galvanometer G shows a deflection which is a constant one if the current is changing at a constant rate. When i is decreasing, the galvanometer deflection is opposite to the one obtained with increasing current.

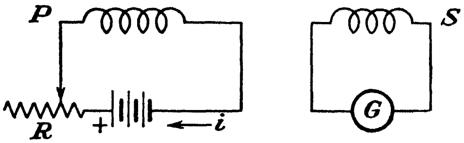


FIG. 24.5. Induction of an emf in the coil S by varying the current in coil P.

Description in Terms of Lines of Force

This variety of circumstances under which electromagnetic induction is observed seems at first overwhelming and confusing. Whereas, paragraphs 1 to 3 make it appear as if it might be possible to arrive at criteria which would allow us to predict the magnitude and the sense of the induced emf from a consideration of the relative motion between pole and coil, observation 4 shows that electromagnetic induction may take place in the absence of any displacement of bodies and thus renders the usefulness of criteria based on the motion of bodies questionable.

All the preceding observations can be described, however, in a uniform language if, instead of considering the motion of the bodies involved in the process of induction, we focus our attention on the magnetic field and on the motion of magnetic-field lines relative to the "secondary" coil S. As we shall show presently, this motion can be imagined to occur whenever the induction of an emf is observed.

Let us consider first the induction of a current in an isolated loop L by a single magnetic pole P (Fig. 24.6), which is the north pole of a bar magnet whose south pole is at an infinite distance from P. Both the loop and the pole are assumed to be located in a vacuum. Let us imagine that the magnetic lines of force issue from P radially like beams of light from a light source which is surrounded by a perforated box through the holes of which thin pencils of light can escape. In position 1 three beams of light (or, rather, three lines of force) are seen to penetrate the loop; in position 2, only one beam is seen to pass through the loop. Thus we see that the change in distance between the pole and the loop, which causes induced currents to flow in L, is associated with a change in the number of magnetic lines of force which penetrate the loop L. We express it by saying that **the induction of an emf in a loop is accompanied by a change in *magnetic flux* through the loop.**

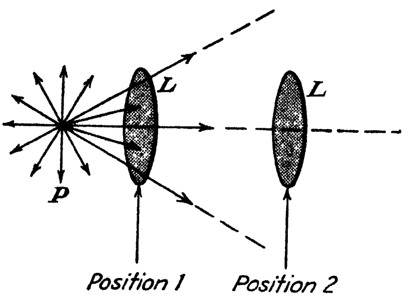


FIG. 24.6. Light-ray model illustrating the change of magnetic flux through the coil L as it is moved from position 1 to position 2.

The other modes of induction can be seen to be similarly accompanied by a change in the flux which penetrates the coil or the loop. If we imagine rays to emanate from the pole N of Fig. 24.2, we can see that the rotation of the loop from position 1 to position 2 reduces the original magnetic flux, which penetrates the coil, to zero. Thus here, too, the process of induction is associated with a change in the magnetic flux through the coil. Similarly, in the case pictured in Fig. 24.3 a diminution in the area of the coil diminishes the magnetic flux through the coil.

Thus we recognize that the change in the magnetic flux which permeates a loop or coil can be correlated with the induction of an emf. This correlation proves particularly fruitful in the considerations of the mode of induction illustrated in Figs. 24.4 and 24.5, where the disposition of the bodies involved remains unchanged. As in previous examples, a change in the number of magnetic-field lines which penetrate the coil S takes place whenever an induced current is observed in S. This can be seen most clearly by considering a modified arrangement in which the coil S is wound over the coil P (Fig. 24.7). In this case all the magnetic-field lines which penetrate P also penetrate S, and anything that will increase the magnetic flux through P will also increase the flux through S. For instance, changing the resistance R *at a uniform rate* will change the current through P and, hence, the magnetic

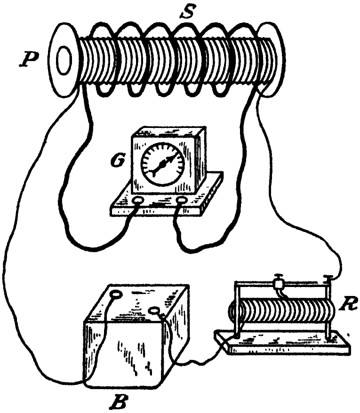


FIG. 24.7. Induction in a secondary coil which is wound over a primary coil.

field intensity and thereby the magnetic flux which penetrates both coils P and S at a *constant rate*. The result will be a *constant* emf induced in S.

If we now insert a long iron bar into coil P, join its ends to form a ring and vary the current exactly as in the previous test, we find that the emf induced in S is μ times as high as in the test conducted in the absence of iron (where μ is the permeability of the iron). The explanation of this effect is provided by the discussion presented in connection with Fig. 23.16. The iron becomes magnetized by the current through the coil P and contributes additional magnetic lines to the flux generated by the current through coil P.

In general, when the coils shown in Fig. 24.7 are submerged in a medium of magnetic permeability μ, the induced emf is increased by a factor of μ. This situation is similar to the relation between the force exerted upon a current and the magnetic field. We have seen in Chap. 23 [equations (23.7*b*) and (23.11*a*)] that this force is proportional to the permeability of the surrounding medium. It is proportional to $B = \mu H$ rather than to the magnetic field strength H. Similarly, experiment shows that **the induced emf is proportional to the "magnetic induction"** $B = \mu H$.

It is convenient to describe the phenomena of electromagnetic induction in terms of the *field of magnetic induction*, which is defined similarly to *magnetic field of force*. In the latter field we associate with every point in space a vector whose magnitude measures the force exerted upon a magnetic unit pole. The local field intensity is represented graphically by drawing H magnetic-field lines per square centimeter (this unit area being perpendicular to $\vec{H}$). *In the "field of magnetic induction" we associate with every point in space a vector whose magnitude* $B = \mu H$ *measures the force exerted at this point upon a unit charge moving with unit velocity at right angles to the field. We represent the intensity of the field of induction graphically by drawing* $B = \mu H$ *field lines per square centimeter. For this reason* B *is also referred to as* **flux density.** The unit of flux density is **1 gauss** corresponding to one field line per square centimeter. The total number of lines of induction passing through an area is referred to as the total *magnetic flux* ϕ.

We shall now summarize the concepts and symbols in terms of which the phenomena of electromagnetic induction are described:

The magnetic flux density (measured in gauss)

$$\boxed{B = \mu H} \tag{24.1}$$

The permeability of the medium μ. (For a vacuum $\mu = 1$, and for air $\mu \cong 1$.)

The magnetic field intensity H (measured in oersteds).

The total magnetic flux

$$\boxed{\phi = A\mu H = AB} \tag{24.2}$$

In this equation A is an area in square centimeters assumed to be perpendicular to the magnetic flux lines and ϕ represents the number of flux lines passing through this area. (A flux of ϕ lines passing through an area is referred to as a flux of ϕ maxwells.)

In all cases of electromagnetic induction the following statement can be shown to be true:

An emf is induced in a closed circuit whenever the magnetic flux penetrating the cross-sectional area of that circuit changes.

Faraday's Law of Induction

We are now ready to examine two questions: (1) How does the magnitude of the induced emf depend on the rate of change of the magnetic flux? (2) How can we predict the sense of the induced emf?

Experiments show that the magnitude of the **emf induced in a closed circuit,** for instance, in a coil, depends only on the number of turns N and on the rate of change of the magnetic flux $\Delta\phi/\Delta t$:

$$\mathcal{E} = -CN\frac{\Delta\phi}{\Delta t}$$

We can define a unit of emf in such a way that C becomes equal to 1. This so-called *electromagnetic unit* (emu) of emf is defined as follows: **1 emu of emf is induced in a loop (N = 1 turn) when the flux through its area varies at the rate of 1 line per second.**

If E is measured in volts, $C = 10^{-8}$ and we get the practical expression†

$$\boxed{\mathcal{E} = -N\frac{\Delta\phi}{\Delta t} \times 10^{-8} \text{ volt}} \tag{24.3}$$

(where $\Delta\phi$ is measured in maxwells, *i.e.*, number of magnetic flux lines).

Equation (24.3) shows that *a small change in flux $\Delta\phi$ may induce a large emf if it occurs during a short enough interval Δt.*

Lenz's Rule

Experience shows further that the direction of the current induced in a loop (or coil) can be determined with the aid of the following rule (Lenz's rule):‡

---

† The reason for the use of the minus sign is explained on p. 400.

‡ This is not the conventional formulation of Lenz's rule.

Imagine the change of flux through the loop to be produced by approaching or removing a magnet pole (as shown in Fig. 24.6) **regardless of how this flux change is actually produced. Then the current induced in the loop makes it act like a magnet of such polarity as to retard the motion of the inducing pole.**

Let us consider the following example to illustrate the use of Lenz's rule: Fig. 24.8 shows a loop in the magnetic field H of the pole N. What is the direction of the induced current which results from an increase in H? Imagine the increase of the field to be accomplished as shown in Fig. 24.8*a*. The downward magnetic flux through the loop L can be increased by moving the pole N toward the loop. According to Lenz's rule the loop should become an electromagnet owing to the induced current and should have such polarity as to repel the approaching north pole; *i.e.*, the north pole of the "loop magnet" should be on the top side of the loop (Fig. 24.8*b*). Now we can ask the question: How must a current flow through the loop in order to make its upper surface act like the north pole of a flat bar magnet? The lines of force inside the loop must be directed upward toward the north pole of the bar magnet. We merely have to draw one typical line of force in order to know the direction of the associated current. We draw first the dotted line of force in Fig. 24.8*b* and affix an arrow to it pointing toward the bar magnet at a point where the line emerges from the loop. We see then that the current giving rise to this line of force must flow into the plane of the paper on the right side of the loop. Thus we find that the induced current is counterclockwise as seen from the top.

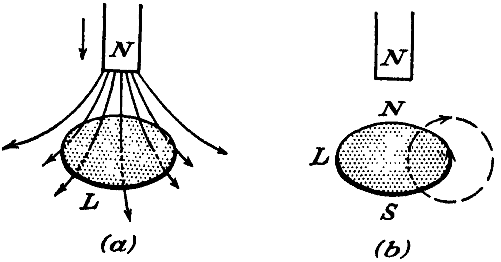


FIG. 24.8.

The same procedure may be used when the magnetic flux penetrating the loop is generated by a coil rather than by a bar magnet, in which case the magnetic flux through L is changed by changing the current in the inducing coil rather than by moving it. We can always imagine the increase in the magnetic flux density at a point to be due to the approach of a magnetic pole, instead of being due to change of current in a coil, and obtain the desired answer by the aforementioned rule.

Since the induced magnetic field in the interior of the loop L (Fig. 24.8*a*) is opposite to the direction of the inducing field, we can also use conveniently the following rule:

The induced magnetic field is such as to counteract the change in the magnetic flux through the loop.

In other words, if the downward magnetic flux penetrating the area of the secondary coil S decreases, the induced magnetic field will be such as to make up for this decrease by adding magnetic flux lines in the interior of the loop to the original lines. Conversely, when the inducing field increases in intensity, the flux lines of the induced field will run opposite to the direction of the inducing field.

Once we establish the direction of the magnetic field induced in the loop, we can easily determine the direction of the induced current which gives rise to this field.

On the Minus Sign in Equation (24.3)

We can now explain the reason for the use of the minus sign in equation (24.3). In this equation $\Delta\phi = \phi_2 - \phi_1 =$ final flux minus the original flux penetrating the secondary coil. For the sake of a concrete picture let us assume that equation (24.3) refers to the arrangement of Fig. 24.7. $\Delta\phi$ is positive when $\phi_2 > \phi_1$, *i.e.*, when the flux generated by coil P is increasing. When shall we consider the emf $\mathcal{E}$ induced in the coil S as positive? We adopt the following convention: $\mathcal{E}$ is to be considered as positive if it gives rise to a current in coil S whose magnetic field has the same direction as the inducing field of coil P (*i.e.*, when the emf $\mathcal{E}$ tends to give rise to a current reinforcing the inducing flux). $\mathcal{E}$ is taken as negative if it generates a current in S which weakens the inducing flux.

Now when ϕ is growing, $\Delta\phi = \phi_2 - \phi_1$ is positive. *We know according to Lenz's rule that the field induced in S counteracts the growth of the inducing flux.* This means that the field induced in S is opposite in direction to the growing field of coil P. According to our convention the emf which gives rise to this opposing field is to be considered as negative. This is accomplished by the minus sign in equation (24.3).

The second possibility involves a diminishing flux in coil P. According to Lenz's rule the current induced in S is such that its magnetic field reinforces the diminishing flux of coil P. The emf induced in coil S should be positive in this case according our convention. And so it is, owing to the use of the minus sign in equation (24.3). Namely, for the diminishing flux $\phi_2 < \phi_1$; hence, $\Delta\phi = \phi_2 - \phi_1$ is negative. The negative sign in front of the fraction $\Delta\phi/\Delta t$ secures a positive value for $\mathcal{E}$.

The Energetic Aspect of Induction

Our first formulation of Lenz's rule indicates that the process of electromagnetic induction is associated with performance of work. This can be clearly seen from the example illustrated by Figs. 24.8*a* and *b*. The loop L in which a current is induced repels the approaching inducing north pole. We have to do work in this process. What happens to the energy we spend? It is converted into kinetic energy of the electrons; *i.e.*, an electric current is established and maintained in the loop. The electrons lose their energy to the atoms of the metal, and the loop gets warm.

Mutual Induction

The arrangement shown in Fig. 24.5 represents a typical circuit to which Lenz's rule and Faraday's law of electromagnetic induction can be applied. According to Faraday's law, the emf induced in coil S is $\mathcal{E}_s = N_s(\Delta\phi_s/\Delta t) \times 10^{-8}$ volt, where N_s is the number of turns of coil S and ϕ_s the magnetic flux through coil S. Now flux ϕ_s is only a portion of the magnetic flux ϕ_p generated by the coil P; only some of the lines of force which pass through coil P (**primary coil**) also pass through coil S (**secondary coil**). But ϕ_s is proportional to ϕ_p; *i.e.*, if we double the current through the coil P, we also double ϕ_p and at the same time ϕ_s.

Thus we can write $\phi_s \propto \phi_p \propto i_p$ or, using a proportionality constant, $\phi_s = Ki_p$. We can now rewrite the expression for $\mathcal{E}_s$ by substituting in it this value of ϕ_s:

$$\mathcal{E}_s = -N_s \frac{\Delta K i_p}{\Delta t} \times 10^{-8} = -10^{-8} N_s K \frac{\Delta i_p}{\Delta t}$$

We now introduce a new constant, the so-called **coefficient of mutual induction** M. By setting $10^{-8}N_sK = M$, we can write

$$\boxed{\mathcal{E}_s = -M \frac{\Delta i_p}{\Delta t}} \qquad (24.4)$$

The value of M is such that when i_p is measured in amperes and Δt in seconds, $\mathcal{E}_s$ is given in volts. *This equation expresses the emf induced in one coil in terms of the rate of change of current in another coil.*

Self-induction

The value of M depends on the distance between the two coils. It increases as the two coils approach each other, so that more of the flux lines which penetrate coil P pass also through the coil S. M reaches a maximum when coil S is pushed over coil P as shown in Fig. 24.7, since the entire magnetic flux generated by coil P passes then through coil S.

The coil S does not necessarily have to be larger than coil P in order to be permeated by the same magnetic flux. We can, for instance, imagine two coils of identical dimensions (solid coil and dotted coil of Fig. 24.9) placed very close to each other, one inside the other as shown in Fig. 24.9. Let us assume that the two coils are very close to each other and yet electrically insulated from each other. The coefficient of mutual induction will have its maximum possible value for these two coils due to the fact that $\phi_s = \phi_p$ in this arrangement. Let us designate by L this maximum possible value of M. Then the emf induced in S is given by

$$\mathcal{E}_s = -L\frac{\Delta i_p}{\Delta t}$$

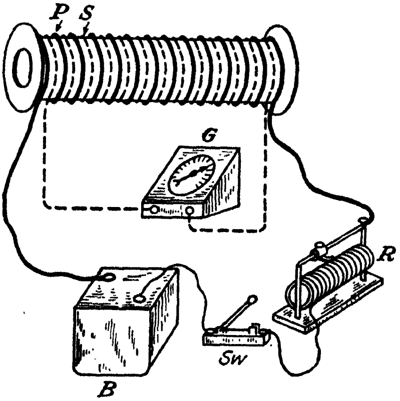


Fig. 24.9. The wires of the two coils P and S are wound next to each other in the fashion of a single "double wire." (The wire of coil S is indicated by a dashed line.)

What would happen now if we press the wires of these two coils against each other so that they lie side by side making contact? The above equation should still hold for the emf induced in coil S due to a variation of the current in coil P effected by manipulating the variable resistance R. But the coil S can hardly be considered now as a separate coil; in fact, coils P and S have been merged into a single coil, and we are led to the strange conclusion that *a variation of current in our new coil* (*SP*)*—which has been created by merging the coils S and P into one—induces an emf in this very same coil of the magnitude*

$$\mathcal{E} = -L\frac{\Delta i}{\Delta t} \qquad (24.5)$$

An appropriate term for this phenomenon would be **self-induction.** *L is referred to as the coefficient of self-induction* or **inductance.**

How could we check experimentally whether or not an emf is actually "self-induced" in a single coil through which a variable current is flowing, giving rise to a changing magnetic flux? Let us suppose that we have a superconductive coil C, *i.e.*, a coil of negligibly small ohmic resistance. We connect it into the circuit shown in Fig. 24.10. When the switch S is closed and a constant current flows through C, the voltmeter V reads zero because the voltage drop $V = iR_c$ vanishes owing to the fact that $R_c \simeq 0$. It is, however, possible to obtain a constant deflection of the voltmeter V by changing the resistance R at a uniform rate. This produces a constant rate of change of the current which flows through the coil C and, hence, a varying magnetic flux, which permeates this coil and changes at a constant rate. As a result of this uniform change in flux, a constant emf is induced in C, which gives rise to a constant reading of the voltmeter connected across C. The voltmeter reading indicates that, *when the current is increasing, the left end of the coil C is positive with respect to the right end. On the other hand, when the current is decreasing, the polarity of the potential difference across the coil is reversed.*

It is now easy to see that *in the case of an increasing current* the coil C acts like a source of emf which opposes the battery B. Whereas the battery tends to circulate a clockwise current in our circuit,† the coil C acts like a source of emf with the positive terminal at the left and the negative terminal at the right end of the coil, tending to produce a counterclockwise current. *When the current is diminishing*, however, the left terminal of the coil becomes negative with respect to its right terminal and its emf aids the emf of the battery: The battery and the coil act like two batteries connected in series, maintaining a clockwise current in the circuit.

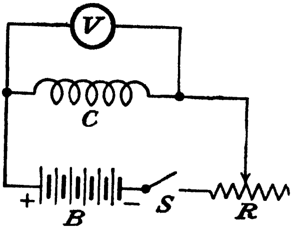


FIG. 24.10.

The voltmeter reading increases in magnitude if the current is varied at a faster rate by a more rapid variation of the resistance R. It can reach very large values (exceeding by far the terminal voltage of the feeding battery B) when the current is changed suddenly by closing or opening the switch S.

The value of the coefficient of self-induction L is increased by any measure which increases the magnetic flux of the coil produced by a given current. Thus, for instance, L can be increased by augmenting the number of turns or by inserting an iron bar into the coil.

The unit of self-inductance is 1 henry. A coil has a self-inductance of 1 henry when an emf of 1 volt is induced in it by a change of current proceeding at the rate of 1 amp/sec. Under these conditions equation (24.5) becomes $1 = -L \times 1$ or $|L| = 1$.

The Energetic Aspect of Self-induction. The following observation which can be made with the setup of Fig. 24.10 appears puzzling at first. When a constant current flows through a coil C which is made of a superconductive material, the potential difference across the coil is $V_c = 0$.

---

† For greater clarity disregard the voltmeter in Fig. 24.10 when reading this section.

This means that the power supplied to the coil C by a constant current is also zero: $P = iV_c = 0$.

But while the current is raised, the potential difference across the coil is not zero and energy is being supplied to the coil at the instantaneous rate of $P = iV_c$ watts. *And yet we observe no heat generation in coil C. How can this observation be reconciled with the law of conservation of energy?*

We observe that the consumption of energy by coil C is accompanied by the creation of the magnetic field of the coil C. *We interpret this observation by ascribing a certain energy content to the generated magnetic field* considering its energy content to be the equivalent of the energy supplied by the battery B to the coil C during the period of the growth of current. Experience shows that there is a constant ratio between the supplied electrical energy and a suitably defined function of the generated magnetic flux. This relationship is similar to the constant ratio between mechanical energy and the amount of heat which appears when work is converted into heat. *We thus interpret this process as a conversion of electrical energy into energy of a magnetic field.*

Can this magnetic energy be converted back into electric energy?

The following experiment gives an affirmative answer: As long as the current remains constant in the setup of Fig. 24.10, $V_c = 0$. Now suppose that we suddenly replace the battery B by a heavy copper wire. Since the battery is eliminated, one might conclude that there is no emf to pump the current i through the circuit. One might, therefore, expect the current to cease instantly, but it does not! It continues to flow despite the absence of a battery, gradually† diminishing in intensity from i to 0. At the same time we notice that (1) the magnetic flux of the coil gradually drops to zero, (2) there is a potential difference across the coil C as long as the current (and, hence, also the magnetic flux) continues to drop. We recognize that energy is being delivered by the coil, which has become the source of an emf of self-induction. Electrical energy is supplied to the resistance R at the rate of i^2R where it is dissipated into heat. The total heat generated is the caloric equivalent of the magnetic-field energy of the coil, which has vanished as soon as the current i has dwindled down to zero.

On account of the large amount of energy which is stored in the magnetic field of a strong electromagnet (for instance, a powerful "eye magnet" for extraction of iron particles), one should turn off the current gradually by increasing the series resistance before switching the magnet off. A sudden interruption of the current (a high value of $\Delta i/\Delta t$) will induce a high emf in the magnet coil. The insulation of the wire is liable to break down, and the large amount of energy released in a short time by the collapse of the magnetic field may seriously damage the electromagnet.

---

† The duration of this process depends on the value of L. It may be in the order of microseconds for small values of L or of minutes for large iron-core coils.

Owing to the large self-inductance of a strong iron-core electromagnet the maximum magnetic field strength is not reached immediately after connecting the magnet to the battery. The waiting time for maximum field strength may be of the order of magnitude of several minutes with very strong electromagnets.

Electromagnetic Induction in an Open Circuit

We have considered so far the induction of an emf in a closed circuit in which the induced emf gives rise to a current. What will happen when we move a straight wire through a magnetic field? Will an emf be induced in this open circuit?

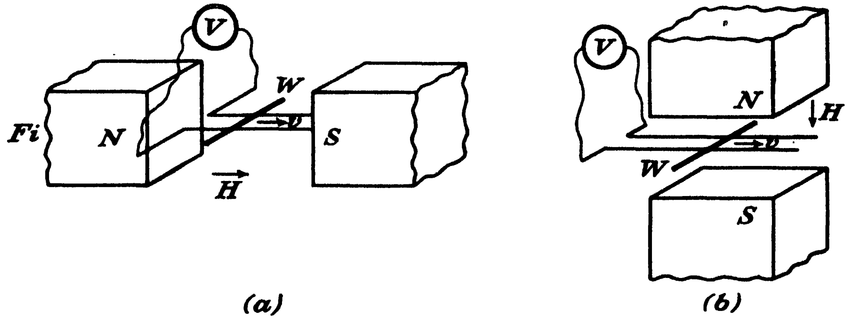


FIG. 24.11*a*. Demonstration of the absence of an induced emf in a wire moving parallel to the magnetic-field lines. *b*. Demonstration of the induction of an emf in a wire moving at right angles to the magnetic field.

The following experiment answers this question: In Fig. 24.11*a* the wire *W* is shown moving in a magnetic field parallel to the direction of the lines of force. As it moves, it makes contact with two rails which can be considered as a continuation of two lead wires of a voltmeter *V*. The voltmeter reads zero, indicating that no emf is induced in the wire in this experiment.

In Fig. 24.11*b* the wire is moving at right angles to the magnetic field. In this case the voltmeter indicates that an emf is induced in the moving wire and furthermore shows that *the induced emf is proportional to the velocity of the wire and to the intensity of the magnetic field.* It can also be shown that *the wire does not have to move precisely at right angles to the field. An emf is induced in it as long as its velocity has a component perpendicular to* $\vec{H}$. The induced emf also depends on the surrounding medium. In a medium of permeability μ the induced emf is μ times as large as in air. In other words, *the induced electromotive force is proportional to $B = \mu H$ rather than to H.*

We can deduce the magnitude of the induced emf from Faraday's law of electromagnetic induction as follows: Let us consider a top-view diagram of the situation pictured in Fig. 24.11*b*. Figure 24.12 shows schematically the wire W and the "rails" connecting it to the voltmeter V. The magnetic lines of flux are perpendicular to the plane of the paper. (The dots indicate the intersections of the magnetic flux lines with the plane of the paper.) When the wire is moved from position 1 to position 2, the cross-sectional area of the closed circuit (which is formed by the hairpin-shaped system of rails plus voltmeter with the wire W placed across the rails) is augmented. The increase in area is $\Delta A = L\,\Delta s$. As the area of the closed circuit is increased, the magnetic flux which penetrates this area is augmented by $\Delta\phi = B\,\Delta A$, where B is the magnetic flux density (in lines per square centimeter).

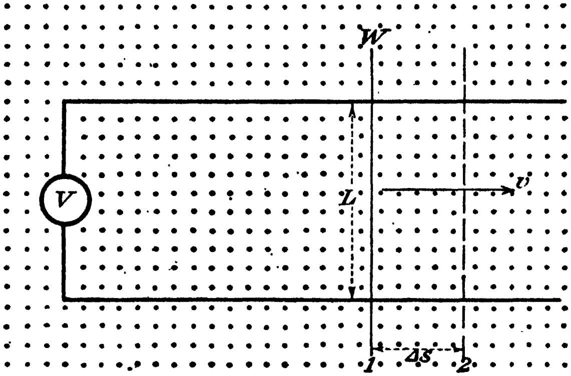


FIG. 24.12. An emf is induced in the wire W as it moves from position 1 to position 2.

According to Faraday's law of induction, this change in flux should give rise to an emf which is given by

$$\mathcal{E} = -\frac{\Delta\phi}{\Delta t} \times 10^{-8} \text{ volt}$$

Using the above-mentioned expressions for $\Delta\phi$ and ΔA we can write

$$\mathcal{E} = -10^{-8}\frac{\Delta\phi}{\Delta t} = -10^{-8}\frac{B\,\Delta A}{\Delta t} = -10^{-8}\frac{B(L\,\Delta s)}{\Delta t}$$

$$= -10^{-8}BL\frac{\Delta s}{\Delta t} = -10^{-8}BLv$$

Thus the absolute value of the emf induced in the section L of the wire W is

$$\boxed{|\mathcal{E}| = BLv \times 10^{-8} \text{ volt}} \qquad (24.6)$$

$B = \mu H$, and Δt is the time during which the transfer of the wire from position 1 to position 2 takes place. $\Delta s/\Delta t = v$ is the speed of the wire.

Equation (24.6) enables us to determine the emf induced in a wire moving in any direction in the magnetic field with the understanding that v is the component of the velocity of the wire at right angles to the magnetic field. The potential difference measured across the length L of the wire by means of a device which does not draw a current (potentiometer) is equal to $\mathcal{E}$.

The object moved across the magnetic field does not have to have the shape of a straight wire. In fact, we can replace the wire in Fig. 24.11*b* by a conductor of any shape without changing the reading of the voltmeter. Equation (24.6) gives the potential difference between two points of a conductor of any shape moving across a magnetic field of flux density $B = \mu H$ provided v measures the velocity at right angles to the field and L measures the distance between two points along a line which is perpendicular to the field and to the velocity.

A Common Description of All Modes of Electromagnetic Induction

The observation of electromagnetic induction in a body due to its motion across a magnetic field introduces a complication in our considerations. We have pointed out previously that electromagnetic induction takes place whenever the magnetic flux through the area of a circuit is changing. Now we see that induction may also occur under seemingly entirely different circumstances. Is it possible to find a common way of describing both modes of electromagnetic induction? Can we show that what happens in a conductor as it moves at right angles to magnetic flux lines is essentially the same process as that which occurs in a closed loop of wire exposed to a changing magnetic flux? Such a unified mode of description is indeed possible. *It is not an explanation of the phenomenon of induction* but merely a convenient and clear method of presentation.

If we imagine the magnetic lines of flux emerging from the pole piece N in Fig. 24.11*b* to be represented in a model by a series of parallel lines directed toward the pole S, the moving wire W will cut these lines as it would have cut material strings in its motion toward the right. **We want to show now that this *imaginary process* of cutting magnetic lines of flux by a conductor can be visualized in all cases of electromagnetic induction.**

Consider, for instance, Figs. 24.6 and 24.8*a*. An emf is induced in the loop by moving a magnetic pole toward it or away from it. If we visualize the lines of flux again as represented by light rays in a model, we see how the number of the lines which penetrate the loop increases as the pole moves closer to it. New lines of flux enter the area of the loop, and we see that in the process of entering the loop the flux lines are cut by the loop in a similar fashion as they are by the moving wire in Fig. 24.11*b*. The same thing happens when the pole is held still and the loop is moved toward it.

It is equally easy to see how the cases of moving the loop away from the pole or rotating the loop in front of the pole or changing the area of the loop in a magnetic field all involve the cutting of magnetic lines of flux by a conductor. The most difficult case to reduce to this mode of description is the mutual induction illustrated in Fig. 24.13. At a particular value of current in the coil P a certain number of flux lines penetrate the area of coil S. When the current i_p is increased, the flux density in the region of the coil S is increased. This means that the magnetic lines of flux must move closer together, moving toward each other as the field intensity is increased. As a result, the magnetic lines can be pictured as moving from the outside into the coil S as the current in coil P is increased. In doing so the magnetic lines of flux will be cut by the turns of the coil S in the same sense as they are cut in the situation represented in Fig. 24.8*a*.

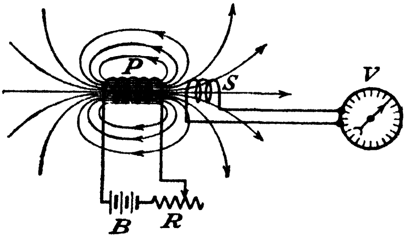


FIG. 24.13.

We have thus succeeded in describing all modes of electromagnetic induction in terms of the *cutting of magnetic lines of flux* by a conductor. We want to emphasize, however, once again that *this is a purely imaginary process. The lines of force are not substantial physical entities like, say, electrons. They are convenient elements of description like the valence bonds of the chemist.* The flux lines enable us to describe and correlate in relatively simple, nonmathematical language a great variety of phenomena which otherwise would have been difficult to coordinate.

The Lorentz Field

Having established the fact that electromagnetic induction occurs whenever there is a relative transverse motion† between a conductor and magnetic lines of flux, we now attempt to determine the mechanism of this phenomenon, *i.e.*, to explain it on the basis of a mechanical theory.

---

† By this we mean that the velocity of the conductor is not parallel to the magnetic field.

We shall begin with a preliminary consideration. We found in Chap. 23 that a charge e, moving in a magnetic field H at right angles to the lines of force with the velocity v, is acted on by a force $f = (\mu/10)Hev = \frac{1}{10}Bev$ dynes [equation (23.11)] which is perpendicular to both $\vec{H}$ and $\vec{v}$. We overlooked at that time a striking feature of this statement. We defined as an electric field a region in which a force is exerted upon an electric charge. But now we are confronted with a strange situation. Imagine that the room in which you are sitting is penetrated by parallel magnetic lines of force running from north to south (Fig. 24.14). Assume that an electron is moving eastward through this room (*i.e.*, away from the reader in Fig. 24.14). This electron will experience an upward force f. *It will experience this force, however, only as long as it is in motion.* When the electron stands still, no forces are acting on it. In other words, *the electrostatic field is zero.* The magnetic compass needle indicates the presence of the magnetic field, *but there is no interaction between the magnetic field and the stationary electric charge.* However, *motion of the charge in any direction except in the direction of the magnetic flux lines makes the charge behave as if it were in an electric field: It is acted on by a force.* We can use a convenient terminology by saying: **The moving charge "sees" an electric field.** Let us refer to this field as the **Lorentz field.** The intensity of the "Lorentz field" is evidently a function of the velocity of the charge. Its magnitude can be determined by comparing the following two equations:

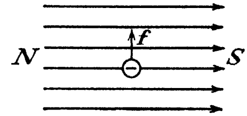


FIG. 24.14.

Force exerted upon a charge e (measured in esu) *in an electric field E:*

$$f = Ee \tag{a}$$

Force exerted upon a moving charge e (measured in abcoulombs, *i.e.*, electromagnetic units) *in a magnetic field:*

$$f = (\mu Hv)e = Bve \tag{23.11a}$$

(where $\vec{H}$, $\vec{f}$, and $\vec{v}$ are mutually perpendicular).

For comparison of equations (a) and (23.11a), we express e in equation (23.11a) in esu. Since the ratio $\frac{1 \text{ emu of charge}}{1 \text{ esu of charge}} = c = 3 \times 10^{10}$ (the magnitude of the velocity of light), the number of emu corresponding to e esu is e/c emu. Hence, if e is the charge in esu, we have to write for equation (23.11a)

$$f = \left(\frac{\mu}{c} Hv\right) e = \frac{1}{c} Bve \tag{23.11b}$$

By comparing equations (a) and (23.11b) we find

$$\boxed{E = \frac{\mu}{c} Hv = \frac{1}{c} Bv} \tag{24.7}$$

for the intensity of the Lorentz field in esu (in dynes per unit charge).

If every moving charge "sees" an electric (Lorentz) field as it speeds through a magnetic field at right angles to the magnetic lines of force, should not the free electrons inside a metal behave like a "trapped cathode ray" and be forced upward when the piece of metal moves through a magnetic field in the same direction as the isolated electron in Fig. 24.14? Would not this upward motion of electrons constitute an induced electric current? How can we check experimentally whether or not such an effect does exist?

Barnett showed the existence of this effect by moving a cylindrical wire through a magnetic field as shown in Fig. 24.11b. This wire consisted of two halves held together end to end. While the wire was moved through the magnetic field, the two halves were separated. It was then shown that the half toward which the electrons were expected to migrate was actually negative, whereas the other half showed an equal positive charge. *This experiment is a close analogy to the demonstration of electrostatic induction* (Fig. 15.1), *i.e.*, separation of charges in an electrical field. *This result can be considered as demonstrating the separation of charges in an electrical Lorentz field which is "seen" by a conductor when moving through a magnetic field.* We thus recognize an analogy between electrostatic induction and electromagnetic induction.

Experiment shows that the potential difference induced in the moving conductor (Fig. 24.11*b*) is the same whether it is the conductor that moves relative to the magnetic field or the magnetic field that moves relative to the conductor. We conclude from such experiments that a Lorentz field depends only on the *relative* velocity between the magnetic field and the charge.

The Mechanism of Electromagnetic Induction

Let us now consider the details of the process of electromagnetic induction. Suppose we move a conductor containing positive and negative electric charges (for instance, a test tube filled with an electrolytic solution) across a magnetic field (Fig. 24.15). The positive and negative ions in the moving test tube will behave like a ray consisting of a mixture of positive and negative ions which, when moving in a vacuum across a magnetic field, would be separated into two distinct beams. Positive and negative ions moving eastward are equivalent to two electric currents: an eastward current (due to the motion of the positive charges) and a westward current (due to the eastward motion of the negative charges). The two currents will experience forces in opposite directions in a magnetic field. In the illustrated case of a solution of NaCl the positive Na ions will be pushed downward and the negative Cl ions upward. How long will this process of separation of charges continue? We see that a negative cloud is being built up near the top of the test tube while a cloud of positive electricity is collecting near the bottom. An electric field is created in the space between these two charged clouds. This field is directed so as to counteract the process of separation of charges in the interior of the test tube; *i.e.*, it tends to move positive charges up and negative charges down. Separation will cease as soon as the electrostatic field between the clouds of Na$^+$ and Cl$^-$ ions has become as strong as the Lorentz field, which is "seen" by the moving ions and which causes the upward migration of the chlorine and the downward migration of the sodium ions.

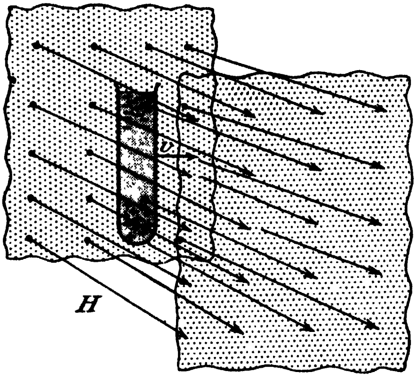


Fig. 24.15. Electromagnetic induction in an electrolyte carried across a magnetic field.

If a similar experiment is performed with a metal rod instead of the test tube, the upper negative cloud is due to an accumulation of electrons while the lower positive cloud corresponds to a region in which there is an electron deficit. As we put the wire *W* of Fig. 24.12 in motion across the magnetic field (the field lines point toward the reader), a transient current should begin to flow through the metal; *i.e.*, the electrons should migrate upward. As soon as the internal electrostatic field has become equal (and opposite) to the Lorentz field $E = (\mu/c)Hv$, the electron current should cease. If we now connect the upper end of the conductor to the lower end through a stationary galvanometer (marked *V* in Fig. 24.12), the instrument will register a current while the electrons flow from the negative to the positive region.

This flow of current tends partly to neutralize the induced positive and negative charges near the ends of the conductor, thus reducing the electrostatic field which opposes the Lorentz field in the interior of the metal. The interior electrostatic "counterfield" is now weaker than the Lorentz field, and a current flows through the conductor as long as its ends are bridged conductively as shown in Fig. 24.12.

The electrostatic field E' must be equal to the potential gradient inside the conductor: $E' = V/L$ (where V is the potential difference between the end points of the portion L of the wire in Fig. 24.12). Since this electrostatic field is equal to the Lorentz field when no current is flowing, we can write

$$E' = \frac{V}{L} = \frac{\mu}{c} Hv$$

or

$$V = \frac{\mu}{c} HLv$$

Since 1 esu of potential difference equals 300 volts (see conversion table in Appendix to Chap. 23), we can express V in volts as follows:

$$V = 300 \frac{\mu}{c} HLv \text{ volts}$$

Now $c = 3 \times 10^{10}$ cm/sec; hence, we obtain

$$V = \frac{300}{3 \times 10^{10}} \mu HLv = 10^{-8} BLv \text{ volt}$$

Thus we have succeeded in deducing the familiar expression for the emf induced in a moving conductor [equation (24.6)] from the electron theory.†

Electromagnetic Induction in a Moving Continuum

The fact that electric charges moving at right angles to a magnetic field experience a deviating force (a "Lorentz force") suggests that an induced emf should be observable in any conductive continuous medium which, as we know, contains freely movable charges which partake in its motion. The following concrete example will illustrate this point:

Imagine a row of parallel wires—separated from each other by a small distance—which are moving behind each other through a magnetic field in the same fashion as the conductor of Fig. 24.15. The same potential difference per unit length will be induced in each of these conductors. If we imagine that they move closer together until they touch one another, we shall have transformed a row of conductors into a continuous moving conducting belt across which a potential difference is maintained by induction (Fig. 24.16).

---

† Although this expression is the same as (24.6), notice that (24.6) was deduced from (24.3), an *empirical* expression based on Faraday's experiments. The deduction presented here is independent of Faraday's work. It can be traced back to (1) Coulomb's law for magnetic poles and (2) Biot's law. The assumption of electrons is convenient, but not necessary, for the presentation.

The emf induced in the belt may be tapped by two sliding electrodes (marked by + and − in the diagram), and the portion of the belt which moves between the poles (N and S) of the magnet can be used as a source of electric power. *This arrangement constitutes a simple d-c (direct-current) generator.*

Figure 24.17 shows a group of separate metal strips all moving with the same velocity into the plane of the paper in a magnetic field. The potential difference induced in each of them varies according to length of the strip:

$$V = BLv \times 10^{-8} \text{ volt}$$

but the potential drop per unit length

$$V/L = Bv \times 10^{-8} \text{ volt/cm}$$

is the same for all of them. Hence, no currents are set up when two of such moving strips are allowed to touch each other.

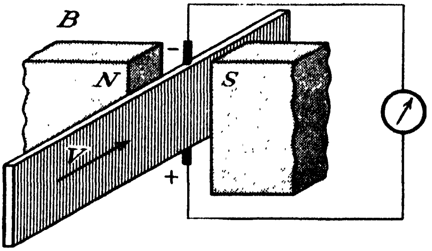


FIG. 24.16.

It is easy to see that a cylindrical conductor moving *in the direction of the cylinder axis* (into the plane of the paper) could be imagined as consisting of adjacent strips as depicted in Fig. 24.17 in cross section. The potential difference induced across its diameter D would then be the same as the potential difference induced across a strip or a wire of length D:†

$$V = BDv \times 10^{-8} \text{ volt} \quad (24.8)$$

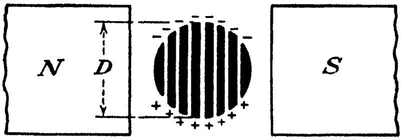


FIG. 24.17. This is the cross section of a cylinder which is perpendicular to the plane of the paper and which has been sliced into the narrow strips shown here. They are moving, in similar fashion to the belt in Fig. 24.16, at right angles to the magnetic field.

If we could measure V with a voltmeter (or, better, with a potentiometer), we could determine the speed of motion v, provided the diameter D and the magnetic flux density B were known.

The Electromagnetic Flowmeter. There is no reason to assume that this cylinder must be solid. One might just as well imagine a cylindrical column of mercury flowing through a pipe (Fig. 24.18). Screws (marked by + and −) can be introduced through the pipe to make contact with the moving liquid. Measurement of the induced emf makes it possible to compute the velocity of the moving fluid if the magnetic field and the pipe diameter are known. Practically, one rather calibrates such a device empirically. The relation between the potential difference between the electrodes V and the average velocity of flow v is a straight line passing through the origin:

---

† This statement is strictly true only in a uniform magnetic field of infinite extent.

$V = (BD \times 10^{-8})v = kv$. We have to establish by calibration only one point on the line, in addition to the origin, in order to determine the linear function $V = kv$.† Our arrangement of Fig. 24.18 thus represents an electromagnetic flowmeter which responds instantaneously to changes in the velocity of flow and has a linear relationship between the scale reading (reading of V) and the rate of flow.

The use of this method of measuring flow, devised by the author, is not limited to a metallic conductor like mercury. We can also use an electrolytic conductor, such as a solution of NaCl, or even as poor a conductor as alcohol! The difficulties which arise with electrolytes due to polarization of the electrodes can be overcome by the use of an alternating magnetic field, which induces an alternating emf in the moving fluid so that unidirectional chemical processes at the electrodes are avoided.

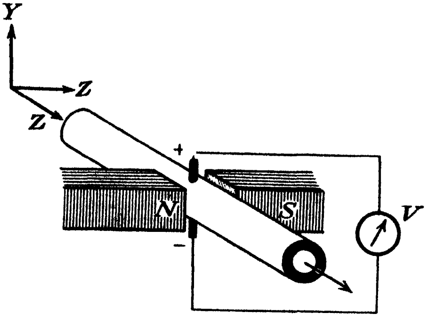


FIG. 24.18. Scheme of the electromagnetic flowmeter.

Electromagnetic Measurement of Blood Flow. We have mentioned in Chap. 12 that one of the chief difficulties in applying a Venturi meter to the measurement of blood flow is due to the fact that the blood vessel has to be cut. This leads to coagulation of blood if no anticoagulant drugs are used or to an abnormal state of the drugged animal if high doses of heparin or of other anticoagulant drugs are injected. In addition, the animal must be anesthetized, which in itself induces a condition widely different from the animal's normal physiological state.

The author showed that the electromagnetic method of measuring flow described above can be used for the measurement of blood flow *without cutting the blood vessel*. The wall of an artery or of a vein is a conductor of electricity. If the blood vessel is placed in a magnetic field as shown in cross section in Fig. 24.19, an emf is induced in the blood as it moves into the plane of the paper. *The blood now acts as an electric power generator* and sends a current through the wall of the artery. This current

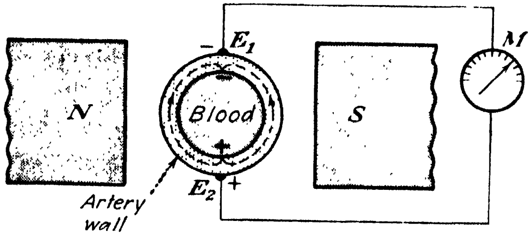


FIG. 24.19. Electromagnetic blood flowmeter. The artery wall is connected to the device M measuring the potential difference between the electrodes E_1 and E_2, which is proportional to the rate of blood flow through the artery.

---

† The calibration consists in allowing a known volume of water to traverse the pipe in a measured amount of time. The voltmeter reading obtained at this rate of flow is noted, and the instrument deflection per unit flow is computed therefrom.

is symbolized in the diagram by two curved arrows inside the artery wall. As a result of this current there is a potential drop between the two points of the artery at which the two electrodes E_1 and E_2 touch it. This potential difference can be recorded by the measuring instrument M The instrument reading is proportional to the current i and thus to the induced emf, which is, in turn, proportional to the instantaneous value of the average velocity of the blood flowing through the cross section of the artery By means of a special recording voltmeter, it is possible to obtain a permanent photographic record of the instrument reading, which represents the instantaneous rate of blood flow as a function of time

Figure 24 20 shows how this method may be applied to an intact artery of an anesthetized animal Figure 24 21 shows the details of the lucite sleeve S which contains the electrodes E_1 and E_2 These electrodes make contact with the wall of the blood vessel which is slipped through the slit Sl into the sleeve

Figure 24 22 shows records of blood flow through the femoral artery of a dog Figure 24 22A shows the marked diminution of the blood flow following the intravenous injection of the drug pituitrin. Figure 24 22B, taken at a higher film speed, shows the details of the variation of the rate of blood flow during the heart cycle We can clearly distinguish between the period of "systolic flow," when the rate of blood flow is high owing to the vigorous contraction of the heart, and the longer period of "diastolic flow," when the blood flows more slowly as a result of the relaxation of the heart This record illustrates the

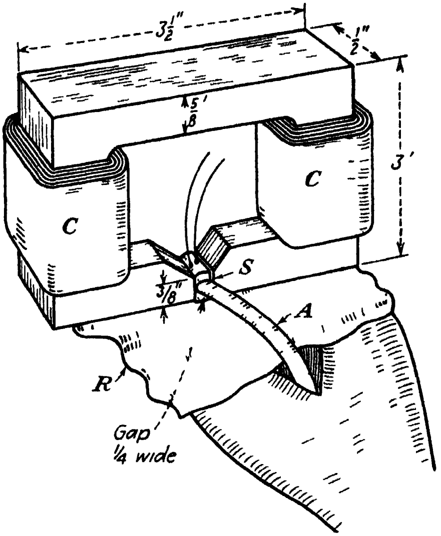


FIG 24 20 Diagram illustrating determination of blood flow in an exposed artery C, field coils of electromagnet, S, lucite sleeve, A, artery, R, dry rubber sheet insulating the exposed section of the artery from the adjacent tissues.

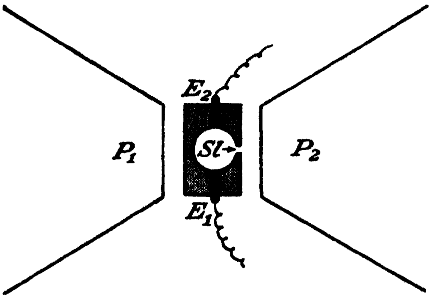


FIG. 24.21. Lateral view of the lucite sleeve Sl, slit, E_1 and E_2, pickup electrodes, P_1 and P_2, the magnet poles

effect of intraarterial injection of 0.07 mg of acetyl-choline into the femoral artery of a 27-kg dog. The segment (*a*) of the record was taken before the injection. The segment (*b*) begins at the moment of the injection. It shows a very marked increase

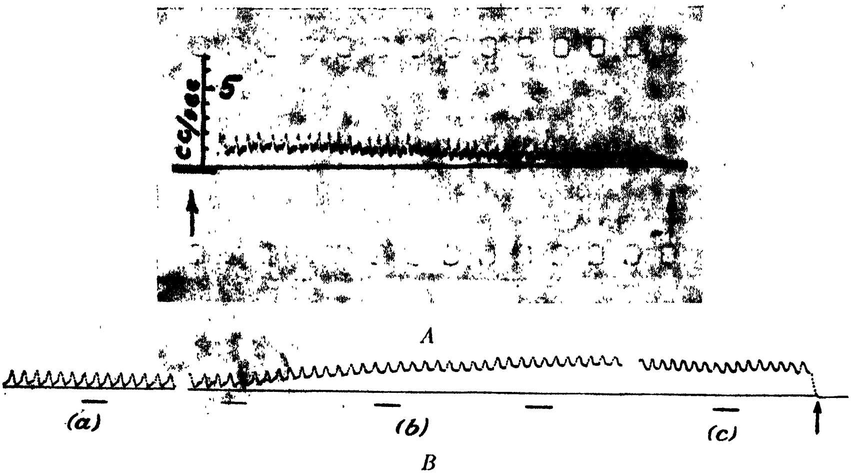


FIG. 24.22*A*. Diminution of blood flow in a dog's femoral artery following intravenous injection of pituitrin. The right arrow indicates the point at which the blood flow was stopped artificially by clamping the artery. Filmspeed 8 cm/min. (*From A. Kolin, Proc. Soc. Expl. Biol. Med., vol.* 46, *p.* 235, 1941.) *B*. Increase in blood flow in a dog's femoral artery following intraarterial injection of acetyl-choline. (*a*) Normal flow before injection. (*b*) The drug is injected at the beginning of the record. (*c*) Stoppage of the blood flow by clamping the artery (indicated by arrow on the right). (*From unpublished records of the author.*)

(*a*) (*b*)

FIG. 24.23*a*. Electromagnetically recorded pumping action of a canulated frog's heart. *b*. The same after a brief immersion of the heart into warm Ringer's solution. (*From A. Kolin, Rev. Sci. Instruments, Vol.* 16, *p.* 109, 1945.)

in the rate of blood flow. In section (*c*) the artery was "clamped off" (arrow) to demonstrate that a zero instrument reading is obtained at zero flow.

The high sensitivity which can be achieved by this method is illustrated by Fig. 24.23, which depicts the pumping action of an isolated canulated frog's heart.

In experiments with dogs it has proved feasible to attach electrodes to a blood vessel, leaving it in its natural place while leading the wires connected to these electrodes to the outside through the skin of the animal. Thus it was possible to record, for instance, the blood flow in the carotid or femoral artery of a dog by placing the entire neck or leg of the animal between the poles of a large magnet after the animal had recovered from the operation in which the electrodes were implanted.

Electromagnetic Induction in a Dielectric. Is electromagnetic induction limited to conductors? What would happen if we allow a dielectric substance, containing electric dipoles, to move through a magnetic field? In an electrostatic field the dipoles would be oriented similarly to the orientation of magnetic needles in a magnetic field. As we have seen before, every charge moving relative to a magnetic field will "see" an electric field if this motion is not parallel to the magnetic-field lines. The moving charge will behave as if it were placed in an electric field (Lorentz field) of intensity $E = (\mu/c)Hv$ (where v is the velocity component perpendicular to the magnetic field). Thus we can predict that the electric dipoles of the dielectric will tend to orient themselves parallel to the Lorentz field. If, for example, our dielectric fills the space between the plates A and B of a condenser which is moving through a magnetic field as shown in Fig. 24.24, the dielectric should become polarized and uncompensated electric charges should appear at the plates A and B as indicated in the diagram. An experiment of this nature was performed by Wilson, and its result confirmed quantitatively the theoretical predictions. We could also perform the converse experiment of leaving the dielectric stationary and polarizing it by moving a magnetic field relative to it.

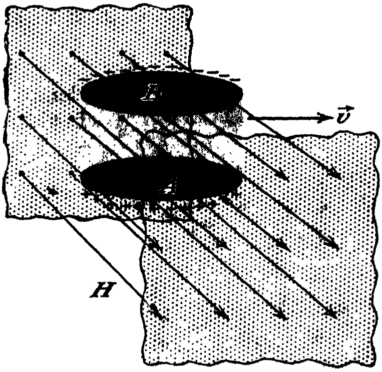


Fig. 24.24. Polarization of a dielectric moving in a magnetic field.

Electromagnetic Induction in a Vacuum. Our ability to polarize a dielectric by moving it relative to a magnetic field or by moving a magnetic field relative to the dielectric shows that the phenomenon of electromagnetic induction does not necessarily involve the presence of a conductor. This suggests the daring question whether the presence of any kind of matter is necessarily required. We know that, when we move a magnetic field past an electric charge located in empty space, the charge will behave as if it were exposed to an electric field of the strength $E = (1/c)Hv$.† We can, if we wish, express these findings by stating that *an electric field is induced in space by a moving magnetic field* and that this electric field can be detected by a charge relative to which the magnetic field is in motion. *The lines of force of the induced electric field $\vec{E}$ are perpendicular to the direction of the magnetic lines and to the velocity vector $\vec{v}$.* The spatial relationship between the vectors E, H, and v is shown in Fig. 24.25.‡

---

† Where v is the velocity component perpendicular to the H vector.

‡ In Fig. 24.25 $\vec{v}$ is the velocity of *the magnetic-field lines* relative to the charge. *The velocity of the charge relative to the field has the opposite direction.* Hence, in a diagram showing the force experienced by a positive charge moving relative to the magnetic field, the force vector has a direction opposite to $\vec{E}$ in Fig. 24.25.

For example, in the experiment illustrated in Fig. 24.13 the magnetic lines of force have been visualized as entering the secondary coil S from the outside as the magnetic field of the primary coil P is increased. The motion of the magnetic lines of flux is toward the center of the coil S. This motion generates an electric field whose lines of force are perpendicular to the magnetic-field lines as well as to the inward radial velocity of motion of the individual magnetic lines of flux. Thus, as a magnetic line cuts across a turn of the coil wire, the electrons in the metal "see" a force which is tangential to the wire; in other words, the induced electric lines of force are represented by concentric circles. This circular electric field affects the electrons in the loops of the coil and whirls them around. If we imagine the secondary coil S of Fig. 24.13 replaced by an empty evacuated doughnut-shaped tube in which electrons are floating around, the electrons will be accelerated in it by the circular electric field and will move in circular or spiral orbits if a centripetal force is provided.

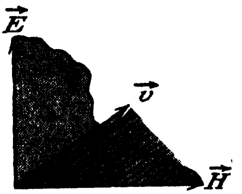


FIG. 24.25.

Principle of the Betatron

The process described above can be visualized more clearly by the diagram of Fig. 24.26. The dots symbolize the intersection of the parallel magnetic flux lines (which point toward the reader) with the plane of the paper. As the field (whose intensity is represented by the number of flux lines per square centimeter) inside the loop increases, we imagine these lines to move radially toward the center of the loop with the velocity v. According to the above discussion, the vector $\vec{E}$ of the induced field intensity will be at right angles to $\vec{v}$ and to $\vec{H}$. Figure 24.26 shows that the E vectors form a circular chain and would tend to accelerate a charged particle continually if that particle could be constrained to move in a circle. By moving many times around a circular orbit, the particle could cover great distances in the accelerating field and acquire a large kinetic energy. As we have seen in Chap. 23 (see Fig. 23.12), it is possible to make an electron describe a circular path by making it move in a magnetic field. A centripetal force is provided by the interaction of the moving charge with the magnetic field. In a constant magnetic field the orbit will be circular if the particle moves in the plane perpendicular to $\vec{H}$ with a constant speed. In the case of a varying magnetic field, the radius of the orbit of a particle will in general tend not to remain constant for two reasons: (1) The radius of

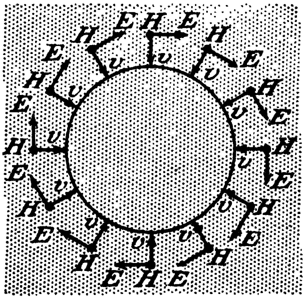


FIG. 24.26.

curvature tends to decrease as H increases; (2) the radius of curvature of the orbit tends to increase as the speed of the particle increases. It is, however, possible to design a magnet in such a way that an increasing magnetic field will accelerate an electron by induction and at the same time will provide an increasing centripetal force varying in such a fashion as to keep the electron on a circular orbit of constant radius.

Let us consider a numerical example illustrating the energy gain of an electron accelerated by induction in a changing magnetic field. Let us suppose that the electron gains 200 electron volts by moving once around in its orbit.† If its velocity were to remain constant at almost the speed of light (3×10^{10} cm/sec)‡ and if the circumference of the orbit were constant and equal to 100 cm, it would take the electron $t = 10^2/(3 \times 10^{10}) = \frac{1}{3} \times 10^{-8}$ sec to complete 1 revolution. If we assume the magnetic field to increase at a uniform rate for the duration of $\frac{1}{300}$ sec, the electron would complete 10^6 revolutions in that time. Thus, the electron would gain $10^6 \times 200 = 2 \times 10^8$ electron volts of kinetic energy.

The sustained acceleration of electrons which describe a circular orbit in a changing magnetic field is the basic idea of the betatron, a new device for generation of exceedingly penetrating X rays and fast electrons for nuclear studies. Electrons of 100 million electron volts have been obtained so far with this method. A betatron is of interest to the medical profession, since it enables one to obtain highly penetrating X rays for therapeutic purposes by means of a relatively compact, easily manageable apparatus.

Deduction of the Laws of Electromagnetism

All the equations stated in Chaps. 23 and 24 have been deduced from two fundamental laws of electromagnetism:

Biot's law:

$$\Delta H = \frac{1}{10} \frac{i\,\Delta L}{R^2} \sin \theta \qquad (23.3)$$

Coulomb's law for magnetic poles:

$$F = \frac{1}{\mu} \frac{p_1 p_2}{R^2} \qquad (14.2)$$

The following is a summary of the basic electromagnetic equations indicating the sequence of their derivation:

---

† 1 *electron volt* is a unit of energy. It is the amount of kinetic energy acquired by an electron which has fallen through a potential difference of 1 volt.

‡ This is true for most of the revolutions.

Coulomb's law

$$\boxed{F = \frac{1}{\mu}\frac{p_1 p_2}{R^2}} \rightarrow$$

Biot's law

$$\leftarrow \boxed{H = \frac{i\,\Delta L}{R^2}\sin\theta}$$

Force upon a conductor carrying a current

$$\boxed{F = BLi} \tag{23.7}$$

$$\downarrow$$

Force upon a moving charge

$$\boxed{F = Bev} \tag{23.11a}$$

$$\downarrow$$

Emf induced in a moving conductor

$$\boxed{\mathcal{E} = BLv} \tag{24.6'}$$

$$\downarrow$$

Emf induced in a closed circuit

$$\boxed{\mathcal{E} = -\frac{\Delta\phi}{\Delta t}} \tag{24.3'}$$

The numerical factors $\frac{1}{10}$ and 10^{-8} have been omitted in these equations, the emf and current being measured in emu. i and v in these equations are the components of current and velocity at right angles to the field.

Equation (24.6) has been derived from (23.11a) at the end of the section entitled Mechanism of Electromagnetic Induction. From equation (24.6) we can derive (24.3) by reversing the reasoning which was used in deriving equation (24.6) from (24.3) with the aid of Fig. 24.12.†

APPENDIX

Principle of the D-c Generator

The d-c (*i.e.*, direct-current) motor illustrated in Fig. 23.25 can be used to generate electric power. Whereas the maintenance of a potential difference between the brushes B_1 and B_2 causes the "armature" (the coil in our simple model) of the motor to rotate, we can conversely generate a nonalternating emf in the coil by rotating it. If the coil is rotated in the direction indicated in Fig. 23.25, the brush B_1 will become the positive and the brush B_2 the negative terminal of the induction generator. As the plane of the coil is rotated from the position shown in the figure to the perpendicular position, the magnetic flux passing through the coil increases from zero to a maximum value. This gives rise to a counterclockwise current in the coil as viewed from the north pole of the generator magnet. As the rotation continues, the number of lines of flux passing through the coil diminishes and becomes zero when the plane of the coil reaches the horizontal position. During this phase of the motion the diminishing flux induces, according to Lenz's rule, a clockwise current in the coil as viewed from the north pole. As a result, the polarity of the terminals of the coil changes, *but the polarity of the brushes does not change,* since contact of the terminal S_2 changes from B_2 to B_1 as the coil passes the position in which the induced current changes direction. Similarly, the contact of S_1 is transferred from B_1 to B_2.

---

† We have deviated in this text from this deductive sequence of presentation in order to illustrate the inductive steps through which the laws of electromagnetic induction have actually been discovered.

This is a crude outline of the principle on which commercial generators of electric d-c power are based.

Diamagnetism

We mentioned in Chap. 15 that there is a class of substances, the so-called *diamagnetic* bodies, which are repelled from a magnetic pole unlike a piece of iron, which is attracted. This was interpreted by assuming that a like pole is induced in the bismuth specimen next to the inducing pole (Fig. 24.27). Since the distance between the like poles is smaller than between the unlike poles, the force of repulsion exceeds the force of attraction and the bismuth sphere Bi is repelled from the magnet pole. How could the theory of molecular magnets explain this fact? It cannot explain it. Dipole molecules should orient themselves as in the case of ferromagnetic bodies. But this, as we have seen, gives rise to attraction.

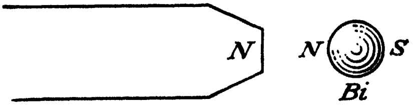


FIG. 24.27. Repulsion of a diamagnetic body.

Can Ampère's theory of molecular currents explain diamagnetism? It seems that it could not any more than could the hypothesis of dipole orientation, since a "molecular current" simply behaves like a magnetic dipole. On second thought, however, we see that this hypothesis suggests a prediction which does not follow from the simple dipole theory. Suppose that we approach a *superconductive* loop to a magnet pole P from position 2 to position 1 in Fig. 24.6. According to Lenz's rule, a current is induced in the loop in such a direction that the loop is repelled from the pole. In other words, the loop is equivalent to a magnet whose north pole is facing the inducing north pole as shown in Fig. 24.27. This experiment suggests a *similarity between diamagnetism and electromagnetic induction.* If we could conceive of the orbits of the atomic electrons, which constitute Ampère's molecular currents, as playing the role of superconductive loops, diamagnetism could be explained as electromagnetic induction taking place in such loops.

Let us consider a concrete example. Suppose that the two loops L shown in Fig. 24.6 are very close together. Assume that they represent orbits of two electrons in an atom, one of the electrons revolving clockwise and the other one counterclockwise. If the two orbits are of equal diameter, and if the electrons move equally fast, the magnetic fields of the two electrons will cancel each other when the atom is out of the range of the magnetic pole P. Let us consider now what should happen to the electron which revolves clockwise as the atom is approached to the pole P, which we assume to be a north pole in this example. It will behave like an electron in a betatron when the field is increasing; *i.e.,* it will be accelerated. The acceleration of the electron is equivalent to an increase of the "loop current" and hence of the associated magnetic field which causes this "loop" to be repelled from P.

At the same time the same change in magnetic field that accelerates the clockwise electron decelerates the counterclockwise one. As a result, the two electrons do not compensate each other any more. The magnetic effect of the clockwise electron, which was accelerated, is now preponderant, and the atom containing both electrons is repelled from the pole P. Thus the electron theory offers a plausible explanation for diamagnetism.

QUESTIONS AND PROBLEMS

1. Two straight vertical wires, each 50 cm long, are moved parallel to each other at right angles to a magnetic field of 2,000 oersteds with a velocity of 250 cm/sec. The two wires lie in a plane which is perpendicular to the direction of motion. The upper ends of the vertical wires are joined by a horizontal wire, and so are the two lower ends. The total resistance of all the wires in this circuit is 10 ohms. How strong is the induced current?

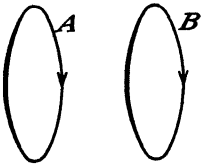
2. Arago observed that a horizontal copper disk held close to a compass needle and rotated rapidly about a vertical axis caused the magnetic needle to rotate in the same sense. How would you explain this observation?

3. A ring, 10 cm inside diam, $\frac{1}{2}$ cm wide, and 0.2 mm thick, is rotated about its axis of cylindrical symmetry. The magnetic field ($H = 1{,}000$ oersteds) is parallel to the axis of rotation, and the frequency of revolution is 100 revolutions per second. How large is the induced emf, and what is the magnitude of the induced current if currents are induced?

★4. How would you wind a resistance coil so as to make self-induction as small as possible?

5. Can you design a brake based on the principle of electromagnetic induction?

6. Two parallel copper loops of equal size face each other at a distance of 1 cm apart. A current of varying intensity and direction flows through loop A. Is there a force of interaction between A and B?



7. Are the following phenomena related: (*a*) the induction of an emf in a conductor which moves in a magnetic field, (*b*) the exertion of a force upon a current which flows through a magnetic field?

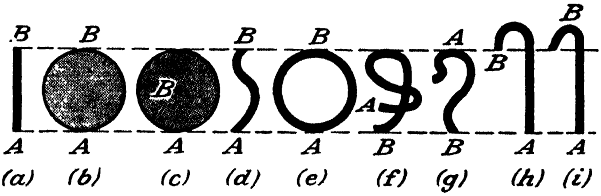
8. If there is a current passing through a solution of $CuSO_4$ and you establish a magnetic field at right angles to the direction of the current, how will the motion of the ions be affected?

9. Can you describe what is going on, according to the electron theory, in a metal sphere as it *accelerates* through a magnetic field in a direction different from that of the field lines?

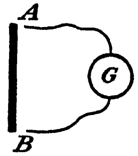
10. How large is the emf induced in a metal wire $L = 10$ cm long which is oriented at right angles to a magnetic field of 1,000 oersteds and is moving with a velocity of 100 cm/sec perpendicularly to the magnetic field?

11. How does the emf induced in a conductor which is moving through a magnetic field depend upon the electrical conductivity of the conductor?

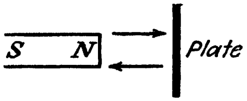
12. The objects shown in the accompanying illustration are moving at right angles to a uniform horizontal magnetic field. The bodies all move with the same velocity away from the reader. How do the magnitudes of the potential differences induced between the points A and B differ for the different bodies?



13. The galvanometer G and the wire AB are moving in the same direction and with the same speed through a magnetic field such that the emf induced in the wire is 0.1 volt. Is it safe to connect the galvanometer terminals to the points A and B if the galvanometer resistance is 100 ohms and if the galvanometer gives a full-scale deflection for a current of 10^{-8} amp?



14. If a copper plate is placed in front of a magnet, which is moved to and fro in front of the plate as indicated by the arrows in the diagram, the plate gets warm. How do you explain this effect?



15. Will the to-and-fro motion of the magnet in Prob. 14 consume just as much energy if the copper plate is replaced by a wooden plate?

16. How would you modify the copper plate so as to avoid or minimize heating, which takes place in the situation of Prob. 14?

17. Can you suggest a detailed design of a setup for electromagnetic measurement of blood flow in an unexposed blood vessel?

★18. Can you suggest an experimental setup to use electromagnetic induction for the study of the velocity distribution in a pipe, *i.e.*, to measure the local velocity of the liquid as a function of the distance from the pipe walls?

★19. Where does the energy imparted to the electrons in the betatron come from?

20. A horizontal zinc rod 50 cm long and 2 cm in diameter is moving with a speed of 100 cm/sec in the direction of its own axis in a vertical magnetic field of 1 oersted intensity. How large is the emf induced in the rod?

21. Determine the direction of the current induced in a horizontal ring when (a) a magnetic north pole is approached to it from the top, (b) when this pole is removed, (*c*) when a magnetic south pole is approached to it from the top, (*d*) when that pole is removed.

22. How large an emf is induced in a coil consisting of 1,000 turns of wire of 0.1 mm diameter wound on a cylinder of 10 cm diameter when a magnetic field of 10,000 oersteds, which penetrates the coil parallel to its axis, is reduced to zero at a uniform rate within 10 sec?

23. A coil of 10,000 turns of thin wire wound on a flat cylinder of 10 cm diameter is placed with its plane perpendicular to the field into a magnetic field of 10,000 oersteds intensity. The two ends of the coil wire are connected to an electrolytic cell consisting of two copper electrodes in a solution of copper sulfate. How much copper is deposited on one of the electrodes when the coil is turned through 180° if the resistance of the cell plus coil is 1,000 ohms?

24. How does the amount of electricity which passes through a closed loop depend upon the time in which the magnetic flux through the loop is changed from ϕ_1 to ϕ_2?

25. How large is the emf induced in a column of blood flowing through an artery of 3 mm internal diameter with a speed of 100 cm/sec at right angles to a magnetic field of 10,000 oersteds?

26. An emf of 0.1 volt is induced in a secondary coil when the current through the adjacent primary coil is diminished at a uniform rate from 20 to 5 amp within 0.5 sec. Find the coefficient of mutual induction between the two coils.

27. When the current through a coil of $0.001\,\Omega$ resistance is changed from 0.1 to 0.01 amp within 10^{-3} sec, a p.d. of 2 volts is recorded across the terminals of the coil. Find its coefficient of self-induction.

28. How large a p.d. will be recorded across the terminals of a "choke coil" (high self-induction) of negligible resistance and an inductance of 40 henrys when the current through the coil is varied at a uniform rate from the value of 1 amp to zero in $\frac{1}{180}$ sec?

29. How large must be an electrostatic field in order to exert upon an electron the same force as a magnetic field of 10,000 oersteds moving at right angles to the field lines with the speed of 300 m/sec relative to the electron?

BIBLIOGRAPHY

Kerst, D., Betatron, in Glasser's "Medical Physics," p. 27, Year Book Publishers Inc., 1944.

Kolin, A., An Electromagnetic Flow Meter: Principle of the Method and Its Application to Blood Flow Measurements, *Proc. Soc. Expl. Biol. Med.*, Vol. 35, pp. 53–56, 1936.

Kolin, A., An Alternating Field Induction Flow Meter of High Sensitivity, *Rev. Sci. Instruments*, Vol. 16, pp. 109–116, 1945.

> We can scarcely avoid the inference that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.
>
> *Maxwell*

PART C: Transition from a Mechanical Model Theory to an Electromagnetic Field Theory

The aim of explaining all phenomena by mechanical model theories was eventually given up. In this section it is shown how the mechanical wave theory of light, after a brilliant initial success, proved inadequate for the phenomenon of polarization of light. This theory of light was originally patterned after the theory of sound. The same role as is played by air in the propagation of sound was supposed to be played by a hypothetical medium, "the ether," in the propagation of light. The theory was given up when it was realized that the properties which had to be ascribed to the ether, in order to explain polarization of light, were mutually inconsistent. The mechanical ether theory of light was replaced by an abstract, nonmechanical "field theory." The periodic process associated with light was no longer imagined as a displacement of ether particles but rather as the variation in space and time of an abstract quantity, the electromagnetic field. The new theory explained, without special assumptions, the polarization of light and, moreover, united the branches of optics and electromagnetism into a single branch of physics. By introducing electromagnetic fields, in which forces are exerted upon charged particles, into optics a bridge has been retained between mechanics and optics.

This section begins with the study of kinematics and dynamics of oscillatory motion (Chap. 25). In Chap. 26 we see how oscillatory motion of particles of matter is propagated in form of mechanical waves. We establish in this chapter the criteria by which the wave nature of a process can be recognized and show (Chap. 27) that sound exhibits all the properties which characterize wave motion.

In Chaps. 28 and 29 we discover many features of similarity between the propagation of sound and of light but also some features in which the two phenomena seem to differ radically. We are faced with an alternative between two hypotheses on the nature of light (wave theory and corpuscular theory) each of which is capable of explaining most of the known facts under suitable auxiliary assumptions. We finally decide in favor of that theory (wave theory) which explains most of the known facts with the least number of special assumptions. This theory proves most successful in explaining known phenomena and in predicting new ones.

In Chap. 30 it is shown how the nonperiodic properties of light have been used to assist our sense of vision and thus to help discover new phenomena and laws of nature. In Chap. 31 we show how the study of the periodic properties of light (wave length and frequency) have enabled us to determine the chemical identity of matter and its state of motion not only in our laboratory but even for bodies which we shall never be able to reach (stars and distant galaxies).

In Chaps. 32 and 33 we study oscillations of electric charges and find that they give rise to nonmechanical waves which can travel in a vacuum. They consist of oscillations of the electromagnetic field in space. These waves have the properties necessary to explain polarization of light. The mechanical ether theory of light is thus dropped and replaced by Maxwell's electromagnetic theory. But historically, the idea of the ether was not dropped immediately; rather it died a slow death. The ether was retained for a while as a sort of a dielectric medium filling all space, until it was recognized that it is not merely superfluous but actually logically untenable (see Chap. 40).

In Chap. 34 evidence for the electromagnetic nature of light waves is discussed, and it is shown how electromagnetic radiations of much shorter wave length (X rays) can be produced.

CHAPTER 25

OSCILLATORY MOTION

a. What kind of motion will result when the force is given by $F = kx$, where k is a constant?

b. Why does the force described by the equation $F = -kx$ lead to a periodic motion?

c. Can you determine the energy content of an oscillating body from its amplitude?

d. What is the purpose of introducing the "reference circle"?

e. What is the significance of the *angular* velocity ω which occurs in formulas of *linear* harmonic motion?

f. What is the phase relationship between displacement, velocity, and acceleration in harmonic motion?

g. At what points does a harmonic oscillator have zero velocity, zero acceleration, maximum velocity, and maximum acceleration?

h. How does a change in the restoring force (for a given displacement) change the frequency of a harmonic oscillator?

i. What factors determine the frequency of vibration of a harmonic oscillator?

j. Under what restrictive condition is formula (25.8) for the period of a simple pendulum valid?

k. What is resonance? Under what conditions does it occur?

Oscillatory Motion

In Part B of this text we have shown how physicists have succeeded in using the laws of mechanics in order to explain thermal, electric, and electromagnetic phenomena. In order to be able to extend the use of the laws of mechanics to phenomena which are not visibly associated with motion of bodies, we had to appeal to our imagination. We assumed the existence of invisibly small particles (molecules, atoms, ions, electrons) moving about in the interior of a body according to the laws of mechanics, which we have established by observing the motion of planets and of smaller bodies in our laboratories. Such hypothetical pictures are designated as "models" of the "physical reality." *Model pictures are to be accepted with caution. They are not meant to commit the physicist to accept a certain mechanical picture as the revelation of the "true design of nature." Their value is chiefly due to the fact that they help us discover mathematical expressions which correlate a set of physical phenomena and which lead us to a correct prediction of new ones. If a model leads to predictions at variance with experience, the physicist is ready to modify it more or less extensively or, if necessary, to give it up.*

Although phenomena of physics could be described in an abstract mathematical fashion without resort to such mechanical models, description in terms of models is widely used. It is considered by many physicists a psychological necessity. For instance, the great physicist Lord Kelvin used to say that he did not feel that he understood a phenomenon until he was able to form a detailed mechanical picture "explaining" it.

We shall see at the end of Part C that mechanical models cannot always be conceived and that tenacious adherence to the mechanical program of physical explanation may actually block progress and lead one astray. It will be shown how the study of a certain mechanical process leads to the establishment of special equations of motion and how these equations may be generalized and applied to processes in which no observable motion is involved. The study of mechanical oscillations and wave motion will provide us with a set of concepts and laws (equations). We shall be able to use these laws for the description of electromagnetic oscillations and electromagnetic waves without adhering to the original mechanical picture. This transfer of the mathematical mode of description to a different field of physics involves an abstraction and generalization of the concepts of oscillations and waves and results in a nonmechanical theory which culminates in Maxwell's electromagnetic theory of light.

Periodic Motion

In our previous studies of motion we have confined ourselves to two simple types of motion: (1) uniform motion ($a = 0$), (2) uniformly accelerated motion ($a =$ constant). Obviously, these are not the only types of motion that occur in nature. In fact, most motions which we observe are, strictly speaking, nonuniformly accelerated. The mathematical treatment of the general case of nonuniform acceleration is quite difficult. There is, however, a special case of nonuniformly accelerated motion, the so-called *simple harmonic motion* which can be handled with relative ease. **We speak of simple harmonic motion in cases where the acceleration of a body changes in direct proportion to the displacement of the body from a certain reference point.** From this statement one might assume the motion in question to be represented by the equation $a = cx$, where a is the acceleration, c a proportionality constant, and x the displacement of the body from a reference point to which we assign the coordinate of $x = 0$. A body moving in accordance with this expression would, however, move faster and faster while receding from the origin with increasing acceleration, evidently acted upon by a force which increases as the distance of the body from the origin grows. This type of motion is practically never encountered in nature and is consequently of no interest to us. However, if we choose a negative proportionality constant ($-c$), we obtain a motion which is of great practical and theoretical interest (**simple harmonic motion**). It is defined by the following equation:

$$\boxed{a = -cx} \tag{25.1}$$

The force responsible for this motion according to Newton's second law ($F = ma$) must be

$$\boxed{F = -(mc)x = -kx} \tag{25.2}$$

The minus sign in equation (25.2) indicates that the direction of the force $\overrightarrow{F}$ is opposite to $\overrightarrow{x}$, the direction of the vector of displacement of the body from the reference point.

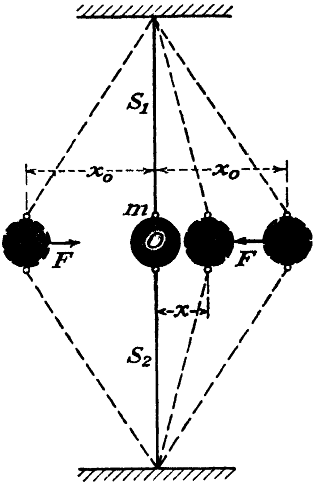


FIG. 25.1. A harmonic oscillator.

We frequently encounter in physics bodies in stable equilibrium, *i.e.*, in positions to which they tend to return after being displaced through a small distance; in other words, there are *"restoring forces" acting toward a point of origin.* [Examples: A weight suspended from a string (pendulum), atoms vibrating in a crystal about their equilibrium positions in the lattice.]

The effect of such a restoring force, acting always toward the point of origin, is to give the body an acceleration toward that point. The force F and, hence, the acceleration a will, in general, vary with the distance x from the origin: $F = f(x)$. If the force were independent of x, we should get a continuous acceleration in one direction and not stable equilibrium. The simplest relation we can think of which will give us the desired result of stable equilibrium is given by equation (25.2). We see from equation (25.1) that the body (mass point) has no acceleration in its equilibrium position ($x = 0$) and has a maximum (restoring) acceleration at the point of maximum displacement x_0.

It can be shown that relation (25.2) holds as a good approximation for sufficiently small displacements in any stable equilibrium position, and hence equation (25.1) holds for any motion in such a region.

The meaning of the above equations can be best understood by considering their application to an actual arrangement in which harmonic motion is observed (Fig. 25.1). A mass m is suspended and held in the position shown in the diagram by means of two elastic strings S_1 and S_2 in such a way that the sideways pull F required to move m through the horizontal distance x from its equilibrium position is proportional to x: $|F| \propto x$. The fact that the restoring force points to the left when the weight is displaced to the right is indicated by a minus sign in the equation: $F = -kx$.

When m is displaced x_0 centimeters to the *right*, a certain amount of potential energy is stored in the distended strings, which exert upon m forces whose horizontal component (F) tends to accelerate m to the *left*. When m is released, it moves with ever-*increasing velocity* toward the central equilibrium position while its *acceleration decreases*. Since $a \propto F$ and $F \propto x$, the acceleration approaches zero as m approaches the equilibrium position $x = 0$ [equation (25.2)]. At $x = 0$ all the original potential energy is present in form of kinetic energy of m. Owing to inertia m moves past O. While its velocity points to the *left*, the restoring force and, hence, the acceleration point toward O, *i.e.*, to the *right*. m is decelerated and finally stops x_0 centimeters to the left of O, which is as far from the equilibrium position as the point from which m was released. Owing to the continued action of F and the resulting acceleration of the mass m, the latter reverses its direction of motion and begins to move to the *right* back toward O, and so on. Thus *a body moving under the influence of a force which is determined by equation* (25.2) *executes a periodic motion.* In the absence of friction, m oscillates between the positions x_0 centimeters to the left and to the right of O. Whereas the displacement x and the velocity v vary their direction in reference to point O, pointing at times toward O and part of the time away from O, *the force and acceleration vectors always point in the direction of O.*

Terminology. We adopt the following terminology for the description of oscillatory processes:

x = displacement.
x_0 = amplitude (maximum displacement).
T = period. (The time which elapses between two successive transitions of m in the same direction through a given point. For instance, the time between two successive transitions through O is not equal to T but only to $T/2$, since m passes through O to the right and then returns, moving to the left. If we wait, however, until m returns to O, moving again to the right, we shall measure a full period T.)
f = frequency.

$$\boxed{f = \frac{1}{T}}$$

(f measures the number of oscillations per second; *i.e.*, f states how many full periods T are contained in the interval of a second. Hence, by definition, $fT = 1$.)

The Reference Circle

Our next step will be to determine the laws of harmonic motion, *i.e.*, to express (1) the displacement, velocity, acceleration, and force as a function of time and (2) the velocity, acceleration, and force as a function of displacement: (1) $x = f_1(t)$; $v = f_2(t)$; $a = f_3(t)$; $F = f_4(t)$; and (2) $v = F_1(x)$; $a = F_2(x)$; $F = F_3(x)$.

These functions can be determined by simple mathematical methods if we consider the relationship between harmonic motion and circular motion. *We shall be able to show that the motion of the suitably projected shadow of an object which moves uniformly in a circular orbit is a simple harmonic motion.* The motion of this shadow can be described conveniently in terms of the circular motion of the body. Consideration of this shadow's motion will enable us to set up the desired equations of motion for a **harmonic oscillator.**

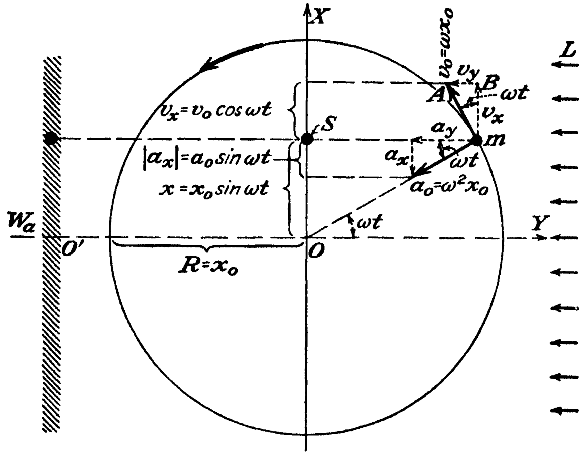


FIG. 25.2.

Let us consider the body m moving with a constant speed v_0 in a circular orbit (Fig. 25.2). The angular velocity $\omega = v_0/R$ is then also constant. Let us assume that sunlight L, whose rays are parallel to the direction of the Y axis, illuminates the wall W_a, which is parallel to the X axis. The shadow S of m will appear on the wall W_a, and as the body m revolves on the circle, its shadow S will move up and down, oscillating about the center point O'. For simplicity, we shall indicate the position of the shadow S along the X axis instead of on the wall W_a. Figure 25.2 is almost self-explanatory. ωt is the angle which has been swept by the radius vector drawn from O to m in t seconds; it is the angle which this radius vector forms with the Y axis. The angle which the tangential velocity $\vec{v_0}$ forms with the direction of the X axis is also equal to ωt, its legs being perpendicular to those of the angle between the line Om and the Y axis.

The displacement x of the shadow S from the equilibrium point O is simply equal to the length of the shadow of the radius $R = x_0$ drawn from O to m. The value of x can be read off from the diagram (triangle OmS):

$$\boxed{x = x_0 \sin \omega t} \tag{25.3}$$

We see from this that x varies with time in a sinusoidal fashion. A plot of x versus time is shown in Fig. 25.3, curve I.

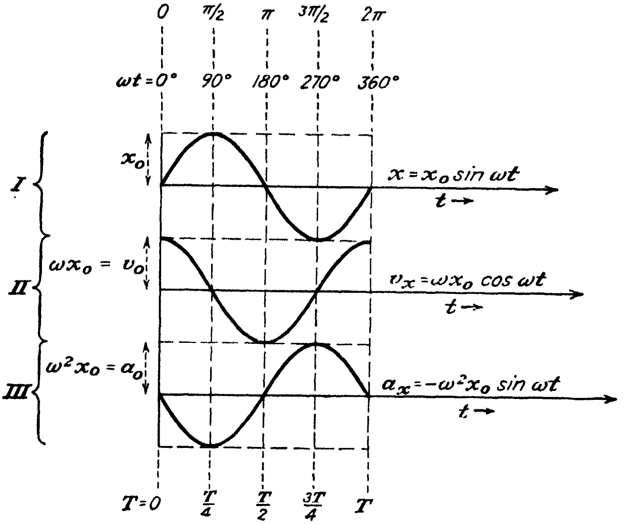


FIG. 25.3. Phase relations between displacement, velocity, and acceleration of a harmonic oscillator. (T is the period of oscillation.)

The velocity of the moving shadow S is found just as easily. The vector $\vec{v_0}$ representing the instantaneous tangential velocity of m can be resolved into two components: v_x parallel to the X axis and v_y parallel to the Y axis. The velocity of the shadow S is given by v_x; the component of the motion of m parallel to the Y axis (v_y) causes no displacement of the shadow's position. The magnitude of v_x is given by the length of the projection (shadow) of the vector $\vec{v_0}$ upon the X axis. We read off from the diagram (triangle AmB)

$$\boxed{v_x = v_0 \cos \omega t} \tag{25.4}$$

Figure 25.3, curve II, shows v_x as a function of time. The velocity of the shadow is seen to fluctuate between the extreme values of $v_x = v_0$ for $\omega t = 0°$ and $v_x = -v_0$ for $\omega t = 180°$. The variations of the velocity v_x can be easily seen by considering the change of length of the shadow of v_0 in Fig. 25.2. When m crosses the Y axis, v_0 is parallel to X and its shadow is longest: The velocity of the shadow S is a maximum ($v_x = v_0 = \omega x_0$) as it passes through the equilibrium point O. From then on the shadow of $\vec{v}_0$ grows shorter until it dwindles down to zero length at the moment when m crosses the X axis. At this moment the vector $\vec{v}_0$ points toward the wall W_a, and its projection has zero length. This means that the body m appears to stand still momentarily at the point where m has reached its maximum distance x_0 from the origin O.

The instantaneous acceleration of the shadow can be determined with equal ease. The body m, owing to its uniform motion in a circular orbit, experiences a constant centripetal acceleration of magnitude $a_0 = \omega^2 R = \omega^2 x_0$ [equation (7.7a)]. The vector $\vec{a}_0$ can be resolved into two components: a_x parallel to the X axis and a_y parallel to the Y axis. The acceleration of the shadow which moves along the X axis is given by a_x. The magnitude of a_x is given by the length of the shadow of a_0 (Fig. 25.2). We read off from the diagram

$$|a_x| = a_0 \sin \omega t \qquad (25.5)$$

and since

$$a_0 = \omega^2 x_0$$

$$|a_x| = \omega^2 x_0 \sin \omega t \qquad (25.5a)$$

We recognize by referring to equation (25.3) that $x_0 \sin \omega t = x$, and hence we can write for equation (25.5a) $|a_x| = \omega^2 x$.

Inspection of Fig. 25.2 shows that the arrow of the *projection* of the $\vec{a}_0$ vector points toward O while the arrow of the shadow of the $\vec{x}_0$ vector (Om line) points away from O. We indicate this fact by the minus sign in equation (25.5b):

$$\boxed{a_x = -\omega^2 x} = -\omega^2 x_0 \sin \omega t \qquad (25.5b)$$

This equation provides the desired proof of the statement that the motion of the shadow S of the revolving body is a simple harmonic motion. It demonstrates (1) that the acceleration of the shadow is proportional to its displacement from a reference point O and (2) that the direction of the shadow's acceleration is opposite to the displacement vector $\vec{x}$. Hence we could, for instance, imagine the following experiment performed: A weight W is suspended on a vertical spring in front of the wall W_a of Fig. 25.2 and is made to oscillate up and down harmonically. We could adjust the frequency of revolution of m so that it equals to the frequency of vibration of W and choose the radius of the orbit of m equal to the amplitude of W. Then if we make the shadow S fall upon the swinging weight W, it will always remain on W, so that both S and W are seen to perform identical motions. *Thus, instead of considering the actual motion of W we can consider the motion of the shadow S of a revolving body m in order to obtain the desired information on the state of motion of an actually oscillating body W.*

The restoring force tending to return the oscillating body to its equilibrium position can be inferred from Newton's second law and equation (25.5*b*):

$$\underline{F = ma} = m(-\omega^2 x) \underline{= -m\omega^2 x} \tag{25.6}$$

Phase Relations

Let us now return to Fig. 25.3. Comparison of curves I and II shows that both displacement and velocity of the harmonic oscillator vary with time in a similar fashion but their maxima and minima do not coincide. **The displacement and velocity are said to be 90° "out of phase."** When the velocity has its maximum in the upward direction, the displacement is $x = 0$. When the upward displacement reaches its maximum value after one quarter of a period, the velocity has dropped to zero; a quarter of a period later, the mass m has returned to zero while the velocity has attained its maximum value in the downward direction. The maxima of velocity are seen to coincide with the zeros of displacement, and vice versa.

The velocity reaches its maximum value earlier than the displacement, *i.e.*, at a smaller value of t; in other words, the maximum of v is to the left of the maximum of x. We express it by saying that there is a phase displacement between v_x and x, the *displacement* x *lagging behind the velocity* v_x.

Comparison of curves I and III of Fig. 25.3 shows that **displacement and acceleration are "in phase opposition."** Both reach their extreme values at the same time. But at the moment when the displacement reaches its maximum eastward value, the acceleration reaches its maximum westward value.

We shall now apply the laws of simple harmonic motion to a few concrete examples in order to clarify their meaning further.

Example 1. A Typical Harmonic Oscillator

A mass $m^* = 245.5$ gm (consisting of lead shot in a container) is attached to a spring, the mass of which is negligible as compared with

m^*. The spring is distended so that the bottom of the mass is on the level of the 5-cm mark of a vertical scale. When the mass is reduced to 245 gm by taking out some lead shot, the spring contracts by $\frac{1}{2}$ cm. When the equilibrium of the new mass of $m = 245.0$ gm is disturbed, it starts to oscillate. Find the period of oscillation.

Solution. From Hooke's law, $F = kx$, we know that a spring distended by x centimeters exerts a restoring force F which increases in proportion to x. The constant ratio $k = F/x$ is called the **spring constant.** *It is numerically equal to the force required to elongate the spring by* 1 *cm.* In our case any deviation of the mass m from its equilibrium position gives rise to a restoring force which tends to move it back to that position with a force that is proportional to the deviation x from the point of equilibrium. Thus the motion is a simple harmonic one, and the following equations should be applicable:

Newton's second law: $F = ma_x$ (*a*)
Spring force: $F = -kx$ (*b*)
From (*a*) and (*b*): $ma_x = -kx$ (*c*)
or $a_x = -\left(\frac{k}{m}\right)x$ (*d*)

The acceleration in harmonic motion is given by

$$a_x = -\omega^2 x \tag{e}$$

By comparing (*d*) and (*e*) we see that

$$\omega^2 = \frac{k}{m} \tag{f}$$

Now ω is the angular velocity of the body which moves around the periphery of the reference circle (Fig. 25.2) with a period T equal to the period of vibration of our oscillating body m. The angular velocity ω can be determined by dividing the angle 2π swept by the radius vector in the course of one revolution by the time T required for one complete revolution. Thus

$$\boxed{\omega = \frac{2\pi}{T}} \tag{25.7}$$

Substituting (25.7) in (*f*) we get

$$\left(\frac{2\pi}{T}\right)^2 = \frac{k}{m} \tag{g}$$

Whence we obtain for the **period of oscillation**

$$\boxed{T = 2\pi\sqrt{\frac{m}{k}}} \qquad (h)$$

To obtain the numerical value of T from (h) we have to find $k = F/x$. We know that removing the weight of $\frac{1}{2}$ gm causes the spring to contract by $\frac{1}{2}$ cm. The change in the distending weight is

$$F = \Delta W = \tfrac{1}{2} \times 980 = 490 \text{ dynes.}$$

Hence,

$$k = \frac{F}{x} = \frac{490}{0.5} = 980 \text{ dynes/cm}$$

Substituting the given data in (h) we obtain for the period of oscillation

$$T = 2\pi\sqrt{\frac{245 \text{ gm}}{980 \text{ dynes/cm}}} = 2\pi\sqrt{\frac{1}{4}}\sqrt{\frac{\text{gm cm}}{\text{gm cm/sec}^2}} = 3.14 \text{ sec}$$

Example 2. The Simple Pendulum

Imagine a heavy sphere of mass m attached to a string. The mass of the string is to be negligible as compared with m, and the diameter of the sphere is to be negligible as compared with the length L of the string (Fig. 25.4).

This arrangement is known as a **simple pendulum.** When the bob m of the pendulum is pulled aside from its equilibrium position, a restoring force is exerted upon it. We shall see that for small values of the angular deviation θ the horizontal restoring force is very nearly proportional to the horizontal deviation x, so that the resulting motion of the bob is a simple harmonic one.

The bob is acted on by two forces: its weight $W = mg$ and the tension of the string P. The combination of the bob and the nearly massless string (mO) can be considered as a body pivoted about the point O. The moment of the force P with respect to O is zero, whereas W has a component $F = W \sin\theta$ at right angles to the line (mO) and hence causes an angular acceleration α of the system. α is determined by the torque $T_R = FL$ according to the fundamental law of rotational motion [equation (8.8)]:

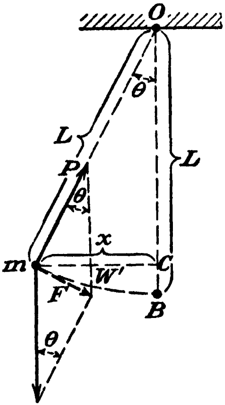


FIG. 25.4. A simple pendulum.

$$T_R = (FL) = I\alpha \qquad (a)$$

where I is the moment of inertia of our body with respect to O. According to equation (8.7),

$$I = \Sigma mr^2 = mL^2 \tag{b}$$

Substituting (*b*) in (*a*) we get

$$FL = (mL^2)\alpha$$

and hence

$$\alpha = \frac{F}{mL} \tag{c}$$

Substituting $F = W \sin\theta = (mg)\sin\theta$ in (*c*) we get

$$\alpha = \frac{(mg)\sin\theta}{mL} = \frac{g}{L}\sin\theta \tag{d}$$

From the triangle *mCO* of Fig. 25.4 we see

$$\sin\theta = \frac{x}{L} \tag{e}$$

hence, substituting (*e*) in (*d*) we obtain

$$\alpha = \frac{g}{L}\frac{x}{L} \tag{f}$$

Now the *tangential* acceleration a of the bob is related to the angular acceleration α through equation (3.5): $a = R\alpha$. In our case $R = L$. Multiplying (*f*) by L we get

$$\underline{a} = L\alpha = L\left(\frac{g}{L}\frac{x}{L}\right) = \underline{\frac{g}{L}x} \tag{g}$$

Since g and L are constants, the above equation states: $a \propto x$. According to our definition, a body moving in a straight line so that its acceleration is proportional to the linear displacement from an equilibrium position (and is opposite to the direction of the displacement) executes a simple harmonic motion. *For very small angles θ ($\theta < 10°$) the motion of the bob can be considered as being nearly linear and horizontal.* The displacement and tangential acceleration become nearly parallel. Hence the above equation (*g*) shows the pendulum motion to be practically a harmonic motion *for small angular amplitudes.*

The period of vibration should be given by the general equation (*h*) derived in the preceding example:

$$T = 2\pi\sqrt{\frac{m}{k}} \tag{h}$$

The value of $k = F/x$ follows from equation (g):

$$k = \frac{F}{x} = \frac{ma}{x} = \frac{m}{x}\left(\frac{g}{L}x\right) = \frac{mg}{L} \qquad (i)$$

Substituting this value of k in equation (h) we get

$$\underline{T} = 2\pi\sqrt{\frac{\cancel{m}}{\cancel{m}g/L}} = \underline{2\pi\sqrt{\frac{L}{g}}} \qquad (25.8)$$

Let us consider a numerical example: How long is the period of a pendulum of length $L = 61.25$ cm?

Answer:

$$T = 2\pi\sqrt{\frac{61.25}{980}} = 2\pi\sqrt{\frac{1}{16}} = \frac{2\pi}{4} = 1.57 \text{ sec}$$

Example 3. Energy of a Harmonic Oscillator

At the moment when the harmonically oscillating mass m passes the equilibrium position, its potential energy is zero; all its energy is kinetic, being determined by its maximum velocity v_0 at that moment. Making use of the relation $v_0 = \omega x_0$ which exists between the maximum velocity v_0 and the angular velocity $\omega = 2\pi/T = 2\pi f$, we can write

$$\underline{E = \frac{m}{2}v_0^2} = \frac{m}{2}\left(\omega x_0\right)^2 = \frac{m}{2}x_0^2\left(\frac{2\pi}{T}\right)^2 = \underline{2\pi^2 x_0^2 f^2 m} \qquad (25.9)$$

We see that *the energy content of the oscillator is proportional to the square of the frequency and to the square of the amplitude.*

Forced Vibrations and Resonance

The frequency with which a freely swinging pendulum bob oscillates is called its **natural frequency,** $f_0 = 1/T = (1/2\pi)\sqrt{k/m}$ [compare with equation (h)]. We can greatly increase the energy of our harmonic oscillator by exposing it to a periodic force which acts on it in the rhythm of its natural frequency. Imagine yourself pushing the pendulum to the right every time it is swinging in that direction of its own accord. Each impulse will increase the kinetic energy of the pendulum. Thus even if the impulse which we impart to the bob is very small, in the course of time the pendulum will have absorbed a large amount of energy. In this way even a small child could make a swing carrying a heavy man oscillate with a great amplitude. This phenomenon is called **resonance.** It plays an eminently important role in many fields of physics and will be frequently referred to in subsequent chapters.

Resonance is not always a desired effect. At times it may lead to disaster. For instance, a platoon of soldiers marching in step may excite vibrations of excessive amplitude in a bridge and wreck it if the frequency of their step happens to coincide with the natural frequency of vibration of the bridge. For this reason the commander always orders soldiers not to march in step when crossing a bridge. In general, resonance is avoided when the frequency of the periodic force does not coincide with the natural frequency of the **resonator** or with a submultiple of that frequency (such as $f_0/2$, $f_0/3$, etc.). In that case there is no fixed phase relation between the acting periodic force and the direction of motion of the resonator, *e.g.*, the pendulum bob in our previous example. For instance, if we start out by pushing the bob to the right when it moves to the right, a few vibrations later the pendulum will be moving to the left when we deliver to it a blow to the right, and instead of increasing its kinetic energy, we shall diminish it. Thus when the periodic force and the pendulum have different frequencies, they will in general be alternately getting out of step and again into step, etc., and the energy of the pendulum will not continue growing as it does in the case of resonance.

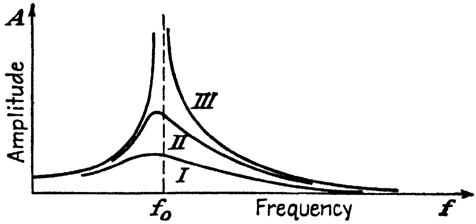


FIG. 25.5. Resonance curves.

Figure 25.5 illustrates how the amplitude of an oscillator changes when it is exposed to a periodic force of varying frequency: The frequency of variation of a periodic force, which is transmitted to a membrane, is plotted along the horizontal axis. Along the vertical axis is plotted the amplitude A of the excited vibration.

Curve I represents the behavior of a vibrator damped by considerable friction. We see how the amplitude increases as the frequency of vibrations is increased and how it assumes a maximum value when the exciting frequency becomes $= f_0$ (*i.e.*, equal to the natural frequency of the membrane). From then on the membrane amplitude drops as the frequency is further increased.

Curves II and III represent the behavior of successively less damped vibrators (less friction). The resonance is said to be "sharper" in the case of lesser damping; *i.e.*, the amplitude excited at nonresonant frequencies is a much smaller fraction of the resonance amplitude. Curve III, representing negligible damping, shows that the amplitude at resonance tends to become infinite in the ideal case of an undamped oscillator.

QUESTIONS AND PROBLEMS

★**1.** A boat, which we consider for simplicity to have the shape of a rectangular box of height a, width b, and length l, is submerged to the depth of $h = 3$ ft. The dimensions are $a = 4$ ft, $b = 4$ ft, and $l = 12$ ft. When its equilibrium is disturbed by lifting it slightly or by pressing it more deeply into the water, it performs a vertical oscillation when released. Answer the following questions:

a. Is the oscillation a simple harmonic one? (Prove your statement mathematically.)

b. What is the period of this oscillation?

c. Does the period of oscillation depend (1) upon the weight of the boat? (2) upon the density of water?

2. A mass of 20 gm distends a spring of negligible mass by 2 cm. When the spring is pulled down 0.5 cm and then released, it performs a harmonic vibration. Find (*a*) the amplitude of this oscillation, (*b*) the period of vibration, (*c*) the energy of this harmonic oscillator. (*d*) Explain why this is a simple harmonic motion. How large a force acts on the mass (*e*) when it is in its lower position, (*f*) when it is in its highest position, (*g*) when it passes the equilibrium point?

3. The bob ($m = 20$ gm) of a simple pendulum vibrates with the frequency of two vibrations per second and the amplitude of 2 cm. Find (*a*) its velocity, (*b*) its acceleration, (*c*) the force exerted upon it in the following positions: (1) at the mid-point of its excursion (*i.e.*, when it is in the lowest position), (2) when its deviation is equal to the amplitude, (3) when its deviation is equal to one-half of the amplitude.

4. A particle of mass $m = 0.2$ gm is attached to the center of a vibrating membrane which oscillates harmonically with a period of $T = 0.001$ sec. It takes a force of 2×10^4 dynes to tear off the particle. If we gradually increase the amplitude of the membrane starting from zero, at what amplitude will the particle fall off?

5. A small sand particle lies on a membrane oscillating harmonically with the frequency of $f = 4{,}000$ vibrations per second. If we increase the amplitude starting from zero, at what amplitude will the sand particle begin to lose contact with the membrane?

6. How large would the acceleration of gravity have to be for a simple pendulum of the period of 1 sec to be exactly 1 m long?

7. How far would we have to get from the surface of the earth in order to reach a point where the length of a simple pendulum of the period of 1 second is 20 cm. ($R_{\text{earth}} = 6.4 \times 10^8$ cm.)

CHAPTER 26

WAVE PROPAGATION

a. What is a wave?

b. Must a wave consist of many crests and troughs?

c. What phase relationship exists between particles of a medium through which a wave is passing?

d. Why do we confine ourselves to the study of sinusoidal waves?

e. How can a complicated vibration be interpreted according to Fourier analysis?

f. What is the fundamental period?

g. What are "harmonics"?

h. What is the distinction between transverse and longitudinal waves?

i. What relationship exists between wave velocity, frequency, and wave length?

j. What is meant by "polarization" of a wave?

k. Why cannot transverse waves exist in fluids?

l. What does Huygens' principle state?

m. How can we predict the future shape and position of a wave front?

n. What is meant by reflection of a wave?

o. What is meant by refraction of a wave?

p. State the law of reflection.

q. State the law of refraction.

r. How is the velocity of wave propagation in solids and fluids influenced by the pressure to which the medium is subjected and by its density?

The Concept of a Wave

When we drop a stone into a pond we see a ring-shaped deformation of the surface, which we call a "wave," spread out in all directions away from the "center of disturbance," *i.e.*, from the point where the stone plunged under the surface. We have the impression that water is moving rapidly away from the center, but a simple experiment demonstrates that this is a mere illusion. Pour a bottle of ink on the surface of the pond before dropping the stone. The colored section of the water will be seen to remain at the same place while the wave traverses it, showing that the propagation of a wave does not involve transport of the *medium.* A piece of cork on the water surface reveals, however, by its upward and downward motion that water is agitated in the region traversed by the wave. A snapshot of the wave reveals that **the wave consists of the propagation of a state of deformation of the liquid surface,** and its inspection explains why the cork is seen to move up and down. Figure 26.1 shows a schematic diagram based on such a snapshot. The water surface is deformed into a craterlike circular mountain consisting of a hill H and a dale *D*, which is depressed below the level of the undisturbed surface. As this wave of deformation passes the cork, the latter is lifted first to the top of the hill, then sinks to the bottom of the dale, and finally comes to rest after the wave has passed it.

It is a general characteristic of wave motion that it does not consist of a transport of matter but rather of the propagation of a state. The following example will show more clearly what we mean by the propagation of a state: Imagine a long column of soldiers (say 2 miles long) marking time to the music of a band which is so loud that all soldiers can hear it. Assume that their performance is such that the right foot of every soldier reaches its highest elevation at the moment when he hears the drum beat. Since sound travels with finite speed (about 330 m/sec), not all of them will raise the right foot at the same time. An observer watching this parade from far away will see that at a certain moment the soldiers in the middle of this column raise their right feet to the highest point at the beat of the drum while those closer to the band will be putting them down, since the sound of the drum has reached them earlier. The feet of the soldiers farther away from the band will be still on their way up when the sound of the drum beat reaches the center soldier. The picture will look about as shown in Fig. 26.2. The right feet of the soldiers will delineate a curve (marked by the dotted line). The apex of this curve will appear to travel away from the band with the speed of sound, since the apex will always coincide with the soldier who hears the drum beat. The motion of these soldiers' feet illustrates a wave motion, the propagation of the state of elevation of a foot.

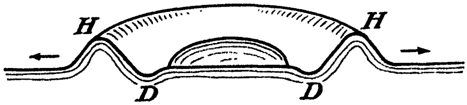


Fig. 26.1. A water surface wave.

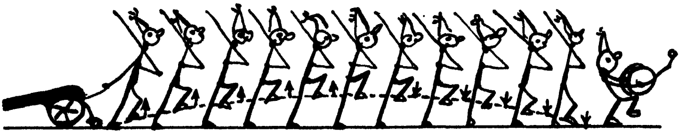


Fig. 26.2.

A useful way of looking at wave motion is illustrated by the following model (Fig. 26.3): Imagine a carpet of coherent beads. Under the carpet is placed a hill-shaped "cam" C, which is moved to the left while the ends of the carpet are kept in place. By looking at the carpet we shall see a deformation travel to the left, reminiscent of a water wave. There will be no motion of beads to the left. They will move only up and down, assuming the elevation which is imposed upon them by the instantaneous position of the cam C.

The concept of a wave can be generalized. Instead of considering the propagation of the state of displacement of particles, we can consider, for instance, the propagation of the state of compression in a gas or of the state of temperature elevation in a solid. We shall then speak of compression waves or of temperature waves, respectively. Discussion of nonmechanical waves will occupy us in later chapters.

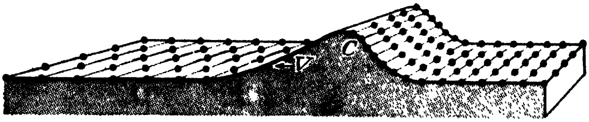


FIG. 26.3. A moving cam model illustrating wave motion.

Fourier Analysis

Figure 26.1 shows a particularly simple type of wave, a **pulse wave.** Not all wave motions we shall encounter are represented by such a simple curve. In fact, the wave shapes encountered in practice (*e.g.*, sound of speech) are usually exceedingly complicated periodic functions such as the curve shown in Fig. 26.4.

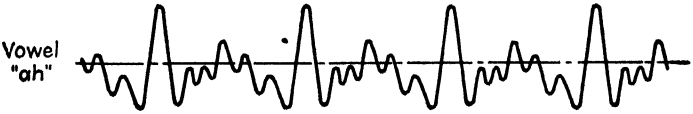


FIG. 26.4. Sound wave corresponding to the vowel "ah." (*From Freeman, "Modern Introductory Physics."*)

One might think that the infinite variety of possible wave shapes would hopelessly complicate the study of wave processes. This is, however, not the case. **Fourier has shown that any curve representing a periodic function, no matter how complicated, can be considered to be the sum† of ordinary sine and/or cosine curves whose reciprocal wave lengths are simple multiples of the reciprocal fundamental wave length of the periodic function itself.‡**

Figure 26.5 gives a concrete illustration of the **Fourier analysis.** Part A shows a box-shaped periodic wave, which we can imagine as extending to infinity to the right and to the left. According to Fourier, this wave can be obtained by adding at every point the ordinates of the following sine waves:

---

† In general, the sum consists of an infinite number of members.

‡ This theorem is not confined to periodic curves; any curve can be represented over a finite range in this manner.

$$y = A_0 \sin x + \frac{A_0}{3} \sin 3x + \frac{A_0}{5} \sin 5x + \frac{A_0}{7} \sin 7x \cdots$$

(A_0, $A_0/3$, $A_0/5$, etc., being the amplitudes of the component sine waves). The component waves are drawn individually in B. If for every value of x we add the values of y of the component curves, we obtain the summation curve C, which has been determined by considering only the first three members on the right-hand side of the preceding equation. This curve roughly approximates the shape of the original curve A. A progressively better approximation is obtained by superimposing an increasing number of successive component sine waves.

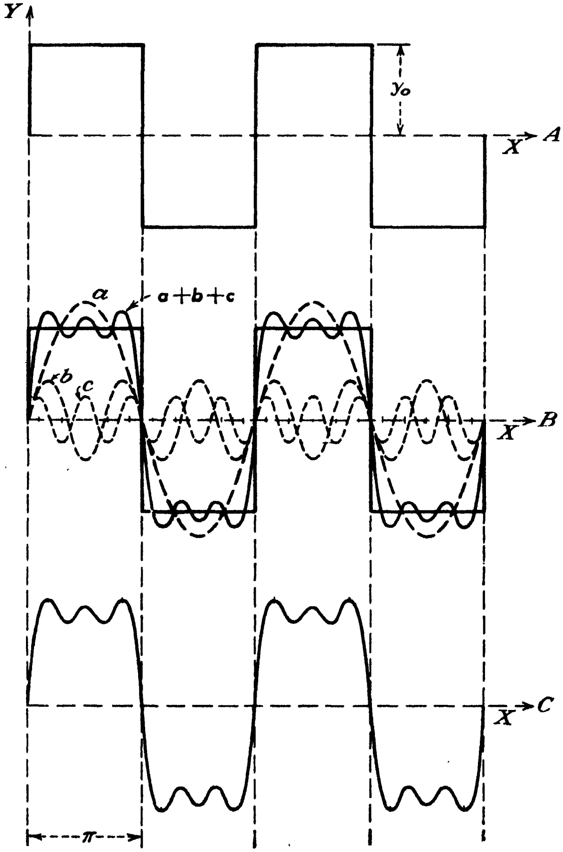


FIG. 26.5. Fourier analysis of a nonsinusoidal wave.

Fourier's theorem is applicable not only to the analysis of the shape of a given wave form but also to the analysis of complex vibrations. If, for instance, the curve A were to represent the instantaneous position of an oscillating body as a function of time (the X axis being the time axis), then the body could be conceived of as performing simultaneously an infinite number of sinusoidal vibrations of different frequencies of which only three component vibrations have been indicated at B in Fig. 26.5. The sinusoidal vibration (a) has the same frequency as the original vibration of Part A. It is called the *fundamental* vibration. The vibration (b) has a frequency three times as high as the fundamental frequency. It is called the third *harmonic.* The vibration (c) has five times as high a frequency as the fundamental and is called the fifth harmonic. There is an infinite number of additional odd harmonics which we might have drawn. But in this particular case they are of diminishing importance with the increasing frequency, since the amplitude decreases as $1/n$ with order n of the harmonic.

Fourier's general expression for any process periodic in time (with a period of repetition $2\pi/\omega$ such as the local displacement in a complex wave) is as follows:

$$Y = f(t) = \frac{A_0}{2} + A_1 \cos \omega t + A_2 \cos 2\omega t + A_3 \cos 3\omega t + \cdots$$
$$+ B_1 \sin \omega t + B_2 \sin 2\omega t + B_3 \sin 3\omega t \cdots \quad (26.1)$$

This expression is called a **Fourier series.** The values of the amplitudes A of the component cosine waves and B of the component sine waves can be determined if the function $Y = f(t)$ is given as a graph. Since ω is given by $\omega = 2\pi/T = 2\pi f$, where T is the period and f the frequency, the above equation can also be written

$$Y = f(t) = \frac{A_0}{2} + A_1 \cos [2\pi(f)t] + A_2 \cos [2\pi(2f)t] + A_3 \cos [2\pi(3f)t]$$
$$+ \cdots + B_1 \sin [2\pi(f)t] + B_2 \sin [2\pi(2f)t] + B_3 \sin [2\pi(3f)t] + \cdots \quad (26.2)$$

We see that the body can be considered as performing, in addition to a vibration at the **fundamental frequency** f, an infinite number of vibrations at **harmonic frequencies** *which are integral multiples of the fundamental frequency* (*2f, 3f, etc.*).

Fourier's analysis is of fundamental importance in the study of vibrations and wave motion. It shows that **we can confine ourselves to the study of sinusoidal processes, since all waves can be represented in terms of sinusoidal waves.**

Transverse and Longitudinal Waves

Let us now design a model of a *medium* in which sinusoidal waves could be propagated. Figure 26.6 shows a row of simple pendulums suspended from very long strings, the bobs being interconnected by springs. Imagine the row of pendulums extended to infinity in both directions of the X axis.

Each pendulum bob experiences a restoring force proportional to the deviation from its equilibrium position when displaced slightly and will consequently execute harmonic oscillations when its equilibrium is disturbed. Imagine, for instance, the bob B pulled toward the reader and then released. It will begin to oscillate, and we shall assume that its initial energy is high enough to enable it to perform several noticeable swings. Why should it swing only a limited number of times if there is no friction? We observe that the oscillation of B agitates the neighboring bobs to which B is coupled by means of springs. The energy, so to speak, "leaks out" of B into its environment. B loses its energy by doing work on its neighbor particles, which now begin to vibrate. Eventually, every pendulum bob in the row vibrates with the same frequency as B but with a different phase, reaching its maximum displacement a little later than bob B. The state of maximum displacement appears to travel along the row of pendulum bobs to each side of B, giving an impression similar to the propagation of the crest of a water wave. If the amplitude of vibration of B is maintained at a constant value, all the other bobs in the row are seen to oscillate with a constant amplitude. If we scan the sequence of pendulum bobs starting from B, we notice that *every bob repeats the motion of its predecessor with a slight delay in time.* Thus, we recognize an essential feature of wave motion: **All particles of the medium through which the wave is propagated vibrate with the same frequency, but there is a progressive change of phase between the particles of the medium (phase lag) as we proceed from particle to particle in the direction of the wave.**

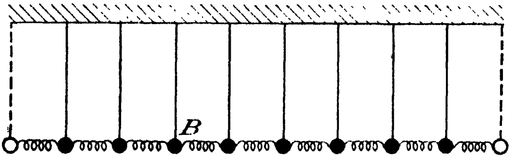


FIG. 26.6. A wave machine for illustration of wave motion.

In the above example the vibration of the particles of the "medium" was perpendicular to the direction of propagation of the wave (the direction of the row of pendulum bobs). A wave in which this relationship exists between the direction of vibration and the direction of wave propagation is called a **transverse wave.**

In the model of Fig. 26.6 it is also possible to move a bob in the direction of the row of pendulums, say to the left. Then pendulum B will vibrate from right to left and so will the other pendulum bobs which are agitated by the vibration of *B*. The line of vibration of the particles of the medium in this case coincides with the direction of wave propagation. Such a wave motion is referred to as a **longitudinal wave.**

The Process of Wave Propagation

Transverse Waves. Let us now consider the details of the mechanism of wave propagation starting with a transverse wave. Imagine a row of particles which are held in their equilibrium position by quasi-elastic restoring forces (Fig. 26.7, top row, black and white circles). The mutual interaction of the particles in this model is effected by springs joining these particles. When particle *A* is disturbed, its motion causes a distortion of the attached spring, which results in the exertion of an accelerating force upon the neighbor particle. The neighbor particle is then also displaced. It distends the following spring and accelerates the

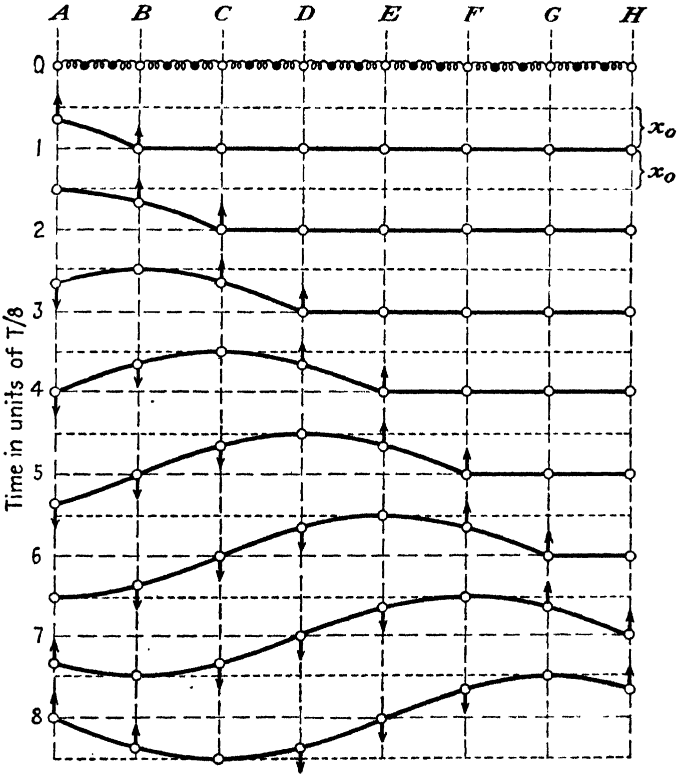


FIG. 26.7. Transverse wave motion.

following particle, etc. Thus the initial disturbance of A is propagated throughout the row.

Let us assume now that particle A vibrates harmonically with a constant amplitude. Let the period of its oscillation be designated by T. To simplify our discussion we shall paint white the particle which is reached after $T/8$ seconds by the disturbance originating at A. We proceed to paint white subsequent particles whose separation is such that the wavelike disturbance requires one-eighth of the period of vibration to travel from one white particle to the next. We thus obtain the landmarks represented by the white circles A, B, C, D, E, . . . , H in the top row of Fig. 26.7.

We shall now examine the mechanism of propagation of a transverse wave by taking snapshots of our row of particles (top row of Fig. 26.7), spacing the snapshots $T/8$ seconds apart. The lines 0 to 8 represent those snapshots.

Snapshot 0. All particles are in the equilibrium position. A has just received an upward impulse and is on its way up.

Snapshot 1. A has moved more than halfway toward its maximum deflection of x_0 during the first time interval of $T/8$ seconds. During the same interval the disturbance has just reached particle B, which now begins moving upward.

Snapshot 2. During the following $T/8$-second interval the disturbance reaches point C. In the meantime A has reached its top position, having been moving upward for one-quarter of a period $[2(T/8)]$, and B has moved as far upward as A is shown to have moved in snapshot 1.

Snapshot 3. A is now on its way down from its position of maximum deviation. It occupies the same position as it did in snapshot 1, but now it is moving in the opposite direction. B has reached the top point and momentarily stands still before the direction of its motion is reversed. C has moved more than halfway up, and D is just about to start moving upward.

Snapshot 4. A is now passing the equilibrium point, moving downward. The apex of the wave, which now becomes recognizable, has advanced from position B (in snapshot 3) to position C.

Subsequent Snapshots. We can see how in the following snapshots the state of maximum deviation is propagated to the right, presenting the typical picture of a traveling wave crest. The point of demarcation between the undisturbed particles and the vibrating ones travels with the same speed to the right as the crest of the wave. In snapshot 7 a well-developed trough of the wave is seen at B (point of maximum downward deviation) and the wave motion has spread over the row of particles from A to H.

If we glance at the sequence of positions of particle A as it is revealed in the snapshots 1 to 8, we see the typical variation of position of a harmonic oscillator. When we examine the positions of particle B down the vertical row, we see that it performs the same motion as A. When we compare, however, the two vertical rows picturing the motion of A and B, we see that *there is a phase lag of one-eighth of a period between the two particles;* thus B assumes at any moment the same position which A occupied $T/8$ seconds ago. The same phase relationship can be established between any two neighboring white particles. *The velocity with which the phase of vibration is propagated to the right is called the* **phase velocity** *of the wave.* We shall refer to it simply as the **velocity of propagation of the wave.**

The Relation $v = f\lambda$. If the row of masses and springs were made longer than shown in Fig. 26.7, we should have obtained the picture shown in Fig. 26.8.

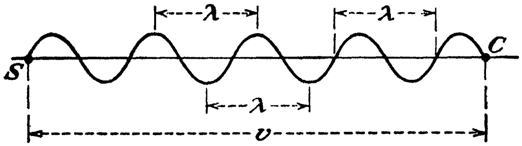


FIG. 26.8.

During the interval of 1 sec the wave motion travels over the distance of v centimeters, where v is by definition the speed of wave propagation. If the source S has performed f vibrations during that second, f wave crests will have emanated from S and will be now equidistantly distributed over the distance v. The distance between the two nearest points of equal phase (*e.g.*, between two neighboring crests or troughs) is designated as the **wave length** λ. We can read off the following relation from our diagram:

$$\boxed{v \doteq f\lambda} \qquad (26.1a)$$

This relation holds for longitudinal as well as for transverse waves. In fact, it can be easily seen that it applies to any conceivable type of wave motion in a homogeneous medium.

Longitudinal Waves. Let us now return to our model of Fig. 26.6. Pulling bob B at right angles to the plane of the paper sets off transverse oscillations of B, which give rise to the propagation of a transverse wave to the right and to the left of B. If, instead of pulling B toward ourselves, we pulled it to the left (remaining in the plane of the paper) and then release it, B would continue oscillating along the line which joins the pendulum bobs, executing a longitudinal vibration which agitates the neighboring bobs, making them vibrate longitudinally and thus giving rise to the propagation of a longitudinal wave issuing from *B*.

The mechanism of propagation of a longitudinal wave is more difficult to visualize than that of a transverse wave. We shall attempt to picture it with the aid of Fig. 26.9.

We begin with the same arrangement of coupled spheres as shown in Fig. 26.7, painting some of the spheres white so that the longitudinal wave requires one-eighth of a period of oscillation to travel from one white sphere to the next white sphere. The row 0 shows the spheres in their equilibrium positions, and the solid vertical black lines indicate this position for the white spheres in all the subsequently shown snapshots (rows 1 to 9). The vertical dotted lines indicate the limits for the maximum deviation of the spheres from their equilibrium position. For instance, the dotted line to the left of the solid line which passes through *C* indicates that the center of the sphere *B* cannot move any farther in the direction toward *C* than this line; in other words, the distance between this line and the solid line through *B* is equal to the amplitude of particle *B*. The subsequent examination of the diagram shows how these dotted lines indicate the amplitude of the longitudinal vibrations.

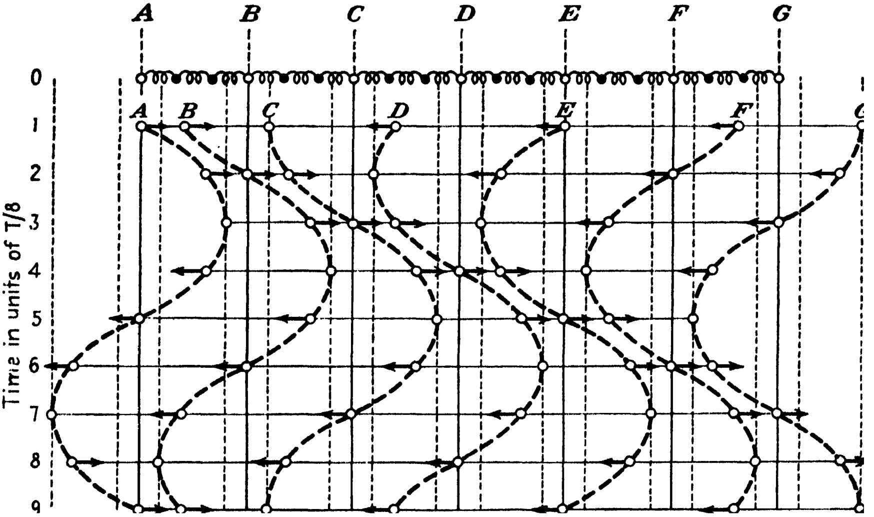


FIG. 26.9. Longitudinal wave motion.

Suppose that *A* is vibrating longitudinally with a constant amplitude and that the entire row of spheres has been set in vibration. As in the case of transverse waves considered above, there will be a phase difference between the vibrations of neighboring particles. We deliberately chose the distance between neighboring white spheres so that the phase lag of a white sphere with reference to its white neighbor to the left is one-eighth of a period. Let us now take snapshots of the row of vibrating spheres (shown in lines 1 to 9), and let us first of all examine the successive positions of the sphere A as they are revealed in the nine snapshots. The successive positions of the sphere A are those to be expected of a harmonic oscillator. Every snapshot indicates the position of the sphere A and its direction of motion. For the sake of clarity all the pictures of A shown in the nine snapshots are joined by a dashed line. Similarly, to avoid confusion of the picture of B with other spheres, all the positions occupied by sphere B in the sequence of snapshots are joined by a dashed wavy line, etc. These dashed lines must not be confused with the contour of a wave. They merely indicate how a given particle changes its position in the course of time.

Motion of A. In snapshot 1, A is seen to traverse the equilibrium point moving to the right. In snapshot 2, it is more than halfway between zero point and position of maximum displacement. In snapshot 3, it has reached its point of maximum displacement and is momentarily at rest. In snapshot 4, it is on its way back passing the same point as in snapshot 2, moving in the opposite direction. The continued picture of its motion is evident from the remaining snapshots.

Motion of B. If the motion of A is indicated in the nine snapshots, how can we determine the motion of B from the motion of A? We know that the vibration of B lags in phase behind the vibration of A by $T/8$. We also know that the time interval between two snapshots is $T/8$ seconds. Owing to the phase lag between A and B the position and velocity of B in snapshot 2 should be identical with those which A possessed $T/8$ seconds previously. The condition of A at that particular moment is shown in the preceding snapshot 1. Hence, the displacement and velocity of B in snapshot 2 is the same as that of A in snapshot 1. Thus we draw the sphere B in the second line as going through the equilibrium point moving to the right. The positions of B in the following snapshots are now easily obtained, since we know how the position of the harmonic oscillator should change in time intervals of $T/8$. We can also deduce what position B occupied in snapshot 1 prior to reaching its position shown in snapshot 2.

Similarly, we can find the displacement and velocity of C in snapshot 2, which are identical with those of B in snapshot 1. The subsequent positions of C are then found like those of B. Thus we obtain the seven dashed lines which indicate the variations of position and velocity of the particles A to G in the course of time.

Let us now examine a typical snapshot, say 3. We see that two of the particles stand still at this moment, namely, A and E (standstill is indicated by lack of arrow). Between these resting spheres there is a region of *condensation.* The spheres B, C, and D are closer together than any of the others. Their distance apart is smaller than in the normal (completely undisturbed) state shown in snapshot 0. On the other hand, the distances between E, F, and G are greater than in the normal state (compare with row 0). The region occupied by these spheres can be designated as a region of *rarefaction.* A region of condensation is sandwiched in between two regions of rarefaction, and conversely, a rarefaction is located between two condensations. In other words, zones of rarefaction and condensation alternate with each other in a longitudinal wave.

We notice that *the direction of motion of the spheres in a condensation coincides with the direction of propagation of the longitudinal wave (to the right in our case), whereas in a rarefaction the spheres move in the direction opposite to that of the progressing wave.*

In snapshot 2 spheres A, B, and C are in a zone of "condensation." In snapshot 3 the condensation is formed by the spheres B, C, D. In snapshot 4 the same can be said of the spheres C, D, E. In snapshot 5 the region of condensation has progressed farther to the right and the spheres D, E, F are closer together than any of the other spheres. We see from this an *essential feature of a longitudinal wave: the propagation of states of condensation followed by states of rarefaction.*

The shift of the region of condensation to the right can be seen clearly by examining the consecutive snapshots shown in Fig. 26.9.

Polarization of Waves

Comparison between the process illustrated in Fig. 26.7 (transverse wave) and the process shown in Fig. 26.9 (longitudinal wave) suggests an important distinction. In Fig. 26.7 the particles are vibrating along the vertical in the plane of the paper while the wave progresses at right angles to the direction of motion of the vibrating particles. This wave is said to be **polarized** in the plane of the paper. We could have vibrated A just as well at right angles to the plane of the paper or at any other angle of inclination to this plane. The particles would have vibrated in a different plane, and the result would have been a "polarized" wave of a different **plane of polarization.**

It is evident that polarization is possible only in the case of transverse waves. In the case of longitudinal waves a "sidedness" of the wave motion as it is displayed by polarized transverse waves is inconceivable. **Longitudinal waves are neither polarized nor polarizable.**

Waves in Fluids and Solids

A longitudinal wave pulse is generated in a medium by creating a local condensation (or rarefaction) of its component particles. The condensation is then propagated as a compressional pulse wave. In our one-dimensional model of a medium of Fig. 26.9, such a pulse is generated by pushing sphere A to the right. Similarly, longitudinal compression waves can be set up in solids or fluids by creating a local condensation of particles by a transient local compression such as an end-on hammer blow against a rod or an explosion in the interior of a liquid or of a gas.

Unlike longitudinal waves, transverse waves cannot exist in fluids. Imagine, for instance, a cylinder suspended in an ideal fluid as shown in Fig. 26.10, executing rotational vibrations about the suspension wire W as a central axis. An ideal fluid is characterized by the absence of friction between two parallel layers of fluid which slide past each other or between the fluid and the surface of a solid which is moving relative to the fluid. Hence, in our case, the fluid will not be affected by the rotation of the cylinder, and no transverse waves will emanate from the latter. If, on the other hand, the cylinder had been embedded in a solid or gelatinous mass, its rotational vibration would have given rise to a transverse wave emanating radially from the cylinder. Actual fluids such as air or water behave approximately as ideal fluids in this respect. **In the interior of liquids and gases transverse waves are practically impossible.**

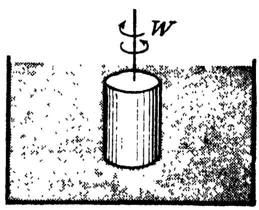


FIG. 26.10.

Velocity of Wave Propagation

Experiments can be performed with the model shown in Fig. 26.7 and 26.9 to demonstrate how the velocity of propagation can be changed by altering the masses of the spheres, the number of spheres per foot, or the elasticity of the springs. An increase of the inertia of the individual spheres or the choice of a denser spacing of the spheres slows down the velocity of the wave. An increase in the elastic constant of the spring (the constant k in the expression for the restoring force $F = kx$), on the other hand, increases the velocity of propagation. Expressions for the velocity of propagation of transverse and longitudinal waves in various types of media can be derived by the application of laws of motion to the displacement of the particles of the medium.

The following consideration illustrates how an expression for the velocity of wave propagation can be deduced from the laws of mechanics. Figure 26.11a shows a transverse pulse wave moving along a string to the right. In Fig. 26.11b a small section of the distorted string near the apex P is considered in detail. The point of the string which is in equilibrium at point O is raised to point P by the passing wave. During the time t the arc AB of the wave moves the distance x to the right. Point A then assumes the position A', which means that the point of the string shown at P in our figure has moved downward into position A', traversing the distance Δy. During the same time the apex P of the wave has moved to position P', which means that the point of the string shown at B has been raised to P'. The time t and the horizontal displacement x of the wave are related by the equation

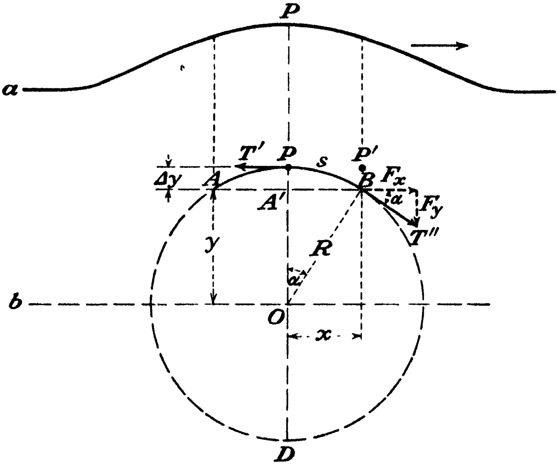


FIG. 26.11.

$$x = vt \tag{26.3}$$

where v is the velocity of propagation of the wave.

The angle α is to be imagined extremely small. In reality, the shape of the very narrow arc AB is approximated by a circular arc of radius R. This arc forms part of the dotted circle in Fig. 26.11*b*. The following well-known geometrical relation holds in the right triangle PBD:† $(A'B)^2 = (PA')(A'D)$. Substituting $A'B = x$, $PA' = \Delta y$, and $A'D = (2R - \Delta y)$ we get

$$x^2 = \Delta y(2R - \Delta y) \tag{26.4}$$

For very small values of α, Δy can be neglected against $2R$. We get then from (26.4)

$$x^2 \simeq \Delta y(2R) \tag{26.4a}$$

Now let us consider the physical aspects of this situation. If we choose α and, hence, Δy small enough, the downward acceleration of the point P can be assumed not to change during the short interval under consideration. Then the displacement Δy will be given by the formula for uniformly accelerated motion:

$$\Delta y = \tfrac{1}{2}at^2 \tag{26.5}$$

---

† The sides PB and BD of this triangle have not been drawn in order not to overload the figure. The triangle PBD is a right one, since its base PD is the diameter of a circle and its apex B lies on the circumference of the circle.

Substituting (26.5) in (26.4a) we get

$$x^2 \simeq \tfrac{1}{2}at^2(2R) = Rat^2 \tag{26.6}$$

Remembering that $x/t = v$ according to equation (26.3) we can write equation (26.6) as follows:

$$\frac{x^2}{t^2} = v^2 = Ra \tag{26.7}$$

or

$$a = \frac{v^2}{R} \tag{26.7a}$$

(Interestingly enough the acceleration of the point P is equal to the centripetal acceleration it would have experienced had it been moving along the arc with velocity v.)

Our last step will be the application of Newton's second law. The arc PB is acted upon by two tangential tension forces T' and T'' which are equal to the tension T in the string. The downward acceleration of the arc PB (which is practically the same as the acceleration of point P for small values of α) must be due to the downward component of T'':

$$F_y = T \sin \alpha = T\frac{x}{R} \tag{26.8}$$

(since $\sin \alpha = x/R$ as can be seen from triangle $A'BO$). According to Newton's second law, $F_y = ma$. Substituting F_y from equation (26.8) the acceleration a from equation (26.7a) and setting $m = \mu s \simeq \mu x$ for the approximate mass of the arc PB, where μ is the mass per unit length of the string,

$$T\frac{\cancel{x}}{\cancel{R}} = (\mu\cancel{x})\frac{v^2}{\cancel{R}} \tag{26.9}$$

hence,

$$\boxed{v = \sqrt{\frac{T}{\mu}}} \tag{26.9a}$$

The velocity of propagation of the transverse wave in a string is given in centimeters per second if the tension T is measured in dynes and the "linear density" μ in grams per centimeter.

Similarly one can derive from the laws of mechanics expressions for the velocity of propagation of longitudinal waves in solids and liquids which will be cited here without proof: **Velocity of propagation of longitudinal waves in a solid:**

$$\boxed{v = \sqrt{\frac{Y}{\rho}}} \tag{26.10}$$

In this equation ρ is the density of the solid, Y is the so-called *Young's modulus*. It is defined as follows: $Y = \dfrac{F/A}{\Delta L/L}$. F is the force of compression or tension in a rod or a wire and A its cross-sectional area in square centimeters. ΔL is the compression or elongation of the solid of length L suffered under the action of the force F.

Velocity of propagation of waves in a liquid:

$$\boxed{v = \sqrt{\frac{B}{\rho}}} \tag{26.11}$$

ρ is again the density of the medium, and B is its *bulk modulus* which is defined as follows: $B = \frac{F/A}{\Delta V/V}$. F/A is the force per unit area, or pressure, to which the liquid is subjected. ΔV is the diminution of volume suffered by the volume V of the liquid when subjected to the pressure F/A.

Velocity of propagation of waves in a gas:

$$\boxed{v = \sqrt{k\frac{p}{\rho}}} \tag{26.12}$$

In equation (26.12) p is the gas pressure and k is the ratio of the specific heat at constant pressure (c_p) to the specific heat at constant volume (c_v):

$$k = \frac{c_p}{c_v}$$

for monatomic gases (such as the inert gases like helium or neon) $k = 1.67$. For diatomic gases (such as nitrogen and oxygen) $k = 1.40$. For the triatomic carbon dioxide $k = 1.30$. To give the student an idea of the order of magnitude of the velocity of sound in various media we cite the following examples at standard conditions of temperature and pressure:

Medium	M/sec	Medium	M/sec
Air	331.8	Alcohol	1,213
Carbon dioxide	258.0	Steel	5,000
Hydrogen	1,269.5	Oakwood	3,850
Water	1,435.0		

Huygens' Principle

Let us now return to our model of a one-dimensional medium of Fig. 26.7 in which waves can be set up and maintained by keeping sphere A in a state of oscillation. Owing to the coupling between A and neighboring spheres, a wave motion is propagated from A. Similarly, one can imagine a plunger moving up and down agitating the surface of water at a point and giving rise to two-dimensional surface waves which issue from the center of disturbance and spread out in concentric circles of growing radius. Every particle of the water surface is agitated as the wave passes its location. Should not an agitated water particle itself act like a miniature liquid plunger, agitating the neighboring fluid particles and giving rise to circular water waves similar to the waves which emanate from the solid plunger referred to above? Huygens concluded that it should. **Huygens postulated that every point in a region of space traversed by waves can be considered as the source of a (secondary) spherical wavelet.** [In the special case of wave propagation in a plane (surface waves) the wavelets are circular.] These small constituent wavelets are a convenient mathematical device to determine the shape of the grand resultant wave in some complicated cases where obstacles or other inhomogeneities in the medium are put into the path of the waves.

Huygens gave the following (approximate) method of constructing the new position of a wave front† from its original position (Fig. 26.12):

Choose arbitrarily a series of points P_1, P_2, P_3, . . . on the original wave front W_1. Draw circles (or spheres in three-dimensional propagation) about them with the radius vt if you want to know the position of the wave front t seconds later (v being the velocity of wave propagation). These small dotted circles are drawn only in part, namely, that part which represents propagation away from the source. They represent the wave fronts of the individual Huygens' wavelets emanating from the selected points of the medium. The resultant wave front is then obtained by drawing the "envelope" W_2, *i.e.*, a common tangent curve to the dotted circles. Since the circumference of the dotted circles represents points of equal phase, the envelope W_2 will unite points of equal phase and will be by definition a wave front.

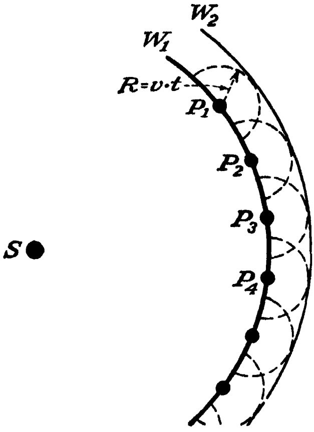


FIG. 26.12. Huygens' construction of a wave front.

The ability of a point of the medium traversed by waves to act as a source of a secondary Huygens' wavelet is demonstrated by the following experiment (Fig. 26.13): A barrier B with a hole H is placed so far from the source S in a pond that waves which arrive at B are not strongly curved (nearly straight wave fronts). On the right side of the barrier B, circular surface waves are seen to emanate from hole H (which is chosen small enough to be considered as a *point source* of waves) in agreement with the prediction of Huygens' principle.

---

† A wave front is defined as a region of equal phase, *i.e.*, a region in which the phase of vibration of the particles of the medium is the same. For instance, the crest of the surface wave in Fig. 26.1 represents a circular wave front.

The significance of Huygens' principle will become clearer when we show how it can be used to derive laws of wave propagation.†

The Law of Reflection. Let us begin with the law of reflection. We can obtain a "beam" of nearly *plane waves, i.e.,* limited wave fronts of negligible curvature, as follows: Place a diaphragm B as shown in Fig. 26.13, but this time with a *large* gap H very far from S, the source of wave motion. The front of the wave which has passed through H will be nearly plane. (The spherical wave front W_2 shown in Fig. 26.13 is observed only when H is small as compared with the relative distance between two consecutive wave crests. If the gap H is many times larger, it simply transmits a portion of the original nearly plane wave without affecting the shape of the wave front except near the edges.) We thus obtain a "beam of plane waves" whose sample wave fronts AB, $A'B'$, and $A''B''$ are shown in Fig. 26.14.

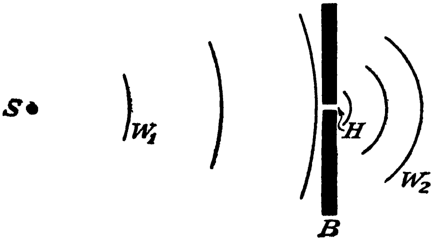


FIG. 26.13.

Figure 26.14 shows a sequence of plane wave fronts (AB, $A'B'$, $A''B''$) [say the wave fronts corresponding to successive maxima (crests)] moving toward the wall R_1R_2 of a medium which does not transmit the waves.‡ What will happen to the wave fronts after they reach the wall? How will this boundary affect the direction of propagation of the incident waves? These questions are answered by Huygens' principle as follows: Consider the first crest AB. We take our first snapshot when the edge A of this crest has reached the boundary R_1R_2. At this moment the particles of medium I located at A become the source of a Huygens' wavelet. t seconds later the edge B of the first crest will have reached the boundary R_1R_2. Point D will now become the source of a Huygens' wavelet crest which, however, at the moment when this second snapshot is taken has still the radius of $R_D = 0$. By this time the wavelet which originated at A has spread over the radial distance $AE = vt$. The solid semicircle indicates the portion of the wavelet which penetrates backward into the medium I, whereas the dotted semicircle beyond the line R_1R_2 indicates where the wave front would be if the medium II transmitted the waves and did so with the same speed as medium I. (The dotted line CD indicates where the wave front AB would have been after t seconds if the boundary R_1R_2 did not exist. The position of this imaginary wave front is obtained by shifting AB parallel to itself through the distance $AC = BD = vt$.)

---

† The complete understanding of Huygens' principle is possible only by considering the phenomenon of "interference of waves" (pp. 514–517). The exhaustive explanation (which is too involved to be taken up in this text) was given by Fresnel long after the principle had been established by Huygens.

‡ It is not necessary to assume that the second medium does not transmit the waves. The phenomenon described below occurs quite generally at any boundary R_1R_2 separating two different media.

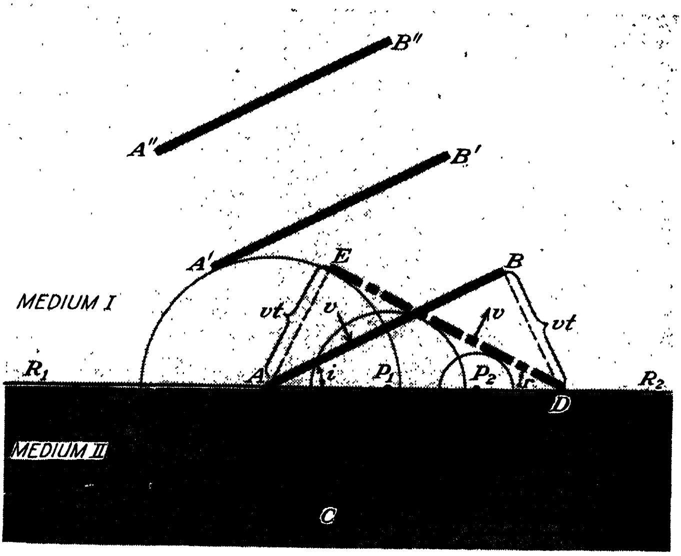


FIG. 26.14. Reflection of plane waves.

Let us now consider in detail the generation of Huygens' wavelets by the wave front AB as it progresses toward the boundary R_1R_2. Imagine that the lines AB and R_1R_2 act like the edges of two intersecting rulers. At first the point of intersection between them is A. The point of intersection becomes the origin of a Huygens' wavelet. As AB moves forward, the point of intersection moves to the right from point A toward point D, reaching D after t seconds. During this time all the points at the boundary between medium I and medium II lying between points A and D become *successively* sources of Huygens' wavelets. Since the wavelets originate at a later time at points which are closer to D, the radii of the Huygens' wavelets issuing from those points are at any time smaller than the radius AE of the wavelet emanating from A. Figure

26.14 shows two typical intermediate points P_1 and P_2 with the corresponding secondary wavelets. Because of the simple proportionality between the distance PD of the origin of a wavelet† from D and the time this wavelet has left to travel until edge B reaches point D, it is easy to see that the wavelets emitted in succession from points P lying between A and D have common tangents ED and CD. The common tangent DE determines the new wave front formed by these wavelets in medium I. This wave front is the front of the so-called **reflected wave.**

Inspection of Fig. 26.14 shows a simple relation between the angle i (**angle of incidence**) between the incident wave front AB and the reflecting surface R_1R_2 and the angle r (**angle of reflection**) formed by the reflected wave front and the surface R_1R_2. We see that

$$\sin i = \frac{BD}{AD}; \qquad \sin r = \frac{AE}{AD}$$

hence,

$$\frac{\sin i}{\sin r} = \frac{BD}{AD}\frac{AD}{AE} = \frac{BD}{AE} = \frac{vt}{vt} = 1$$

hence,

$$i = r$$

This is the law of reflection. The angle of incidence is equal to the angle of reflection.

Law of Refraction. What will happen to the wave front AB when it penetrates into a medium II in which the wave travels, say, more slowly than in medium I?

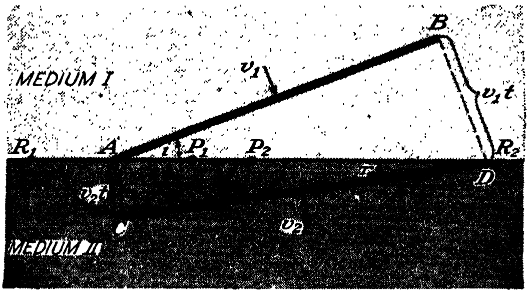


FIG. 26.15. Refraction of plane waves.

Let us consider the wave front AB in Fig. 26.15 at the moment when its edge A reaches the boundary R_1R_2. Point A becomes the source of a secondary wavelet. (We shall draw only that part of the secondary wavelets which penetrates into medium II, the part which is propagated in medium I being of no interest to us in this consideration.)

---

† P being also the intersection of the lines AB and R_1R_2.

t seconds after the edge A reaches the boundary R_1R_2, the edge B reaches point D. By this time the entire wave front has entered the medium II. What is the direction of the wave front in the second medium? We obtain the wave front in medium II as the envelope to the Huygens' wavelets emanating from the points at the boundary which are induced to emit secondary wavelets in succession as the point of intersection between the lines AB and R_1R_2 sweeps from A toward B. Three typical wavelets emanating from the points A, P_1, and P_2 have been drawn as they would appear t seconds after the edge A of the wave front AB reached the boundary. They travel in the medium II with the speed v_2. The direction of the wave front in medium II is seen to be different from its direction in medium I. The wave is said to have been **refracted.** It now proceeds in a different direction, which is perpendicular to the orientation of the new wave front CD.

The following simple relationship can be read off from the diagram:

$$\sin i = \frac{v_1t}{AD} \qquad \text{and} \qquad \sin r = \frac{v_2t}{AD}$$

hence,

$$\frac{\sin i}{\sin r} = \frac{v_1t}{AD}\,\frac{AD}{v_2t} = \frac{v_1}{v_2} = n \tag{26.13}$$

This is the **law of refraction.** It describes how the direction of the wave front changes as it passes from a medium in which the velocity of wave propagation is v_1 to a medium in which the wave velocity is v_2. The ratio $n = v_1/v_2$ is the so-called **index of refraction** or, more precisely, the *relative index of refraction of medium* II *with respect to medium* I.

We see that, if $v_2 < v_1$, the wave front in the second medium is more nearly parallel to R_1R_2 than it is in the first medium. (Draw the refracted wave front for the case $v_2 > v_1$.)

QUESTIONS AND PROBLEMS

1. Can you visualize wave motion which is a combination of longitudinal and transverse wave motion? Give an example. How does a cork on the wavy surface of a lake move?

2. Why does throwing a stone into water generate a wave?

3. How could you generate "heat waves" in a medium?

4. How would you modify the parts of the model shown in Fig. 26.6 in order to increase the velocity of waves which it demonstrates?

★5. What would happen to the contour of a complicated nonsinusoidal wave if the velocity of wave propagation in the medium were dependent on the frequency of vibration?

6. A source vibrating at constant frequency sends out waves which pass in succession through two media in which the velocities of propagation are v_1 and v_2, respectively ($v_1 > v_2$). Will the detected frequency be affected as the waves pass from medium I to medium II? Will the wave length be affected?

7. What is the direction of motion of the particles in the region of condensation in a compression wave? and in the region of rarefaction?

8. How could you generate (*a*) transverse, (*b*) longitudinal waves in a solid? Give examples.

9. A wooden board is thrown into a pond. Construct the resulting wave front.

10. It has been observed that only longitudinal earthquake waves are transmitted from the "focus" of an earthquake to an observation post in those cases where the wave trains travel through the deep interior of the terrestrial globe. Quake waves which travel near the surface have been observed to consist of both longitudinal and transverse waves. What conclusions could you draw from these observations about the state of matter in the central region of the earth?

11. Determine the velocity of sound waves in air at atmospheric pressure and 0°C.

12. Find the wave length of a sound wave of (*a*) the lowest frequency to which the human ear responds: $f = 16$ vibrations per second, (*b*) the highest frequency to which the adult human ear responds: $f = 20{,}000$ cycles/sec.

13. Can you conclude from the fact that temperature does not occur in equation (26.12) that the velocity of sound in air is independent of temperature?

14. Compute the per cent change of the wave length of a sound wave of frequency $f = 1{,}000$ vibrations per second which occurs when the atmospheric pressure rises from 75 to 76.5 cm Hg.

15. Predict how the velocity of propagation of waves in a gas should change with gas pressure at constant temperature.

16. A log 5 ft long and 0.5 by 0.5 ft in cross section is dropped horizontally into a lake. Determine by graphical construction how the wave front issuing from the log changes in the course of time. Draw several successive wave fronts. What is the shape of the wave front very far away from the log?

17. The velocities of wave propagation in two gases separated by a thin plane membrane are $v_1 = 400$ m/sec and $v_2 = 500$ m/sec. Determine the direction of propagation of the wave after passing through the membrane under the following conditions: (*a*) The incoming wave fronts form an angle of 30° with the plane of the membrane and move in the direction from gas 1 to gas 2; (*b*) the conditions are the same as in (*a*) except that the waves move in the direction from gas 2 toward gas 1. Determine the index of refraction of gas 1 relative to gas 2.

18. A wave of wave length $\lambda = 5$ cm travels along a string when it is agitated at a point with a frequency $f = 200$ vibrations per second. The mass per unit length of the string is $\mu = 0.01$ gm/cm. How large is the tension in the string?

19. The velocity of sound in water at 0°C and atmospheric pressure is $v = 1{,}435$ m/sec.

a. Find the bulk modulus of water.

b. Find the change in the volume of 1 liter of water obtained when the outside pressure is raised from 1 to 10 atm.

20. A plane explosion wave generated under the surface of a lake moves upward 30° off the vertical. In what direction will it proceed in air? (Assume the following wave velocities: for air $v' = 330$ m/sec, for water $v'' = 1{,}435$ m/sec.)

CHAPTER 27

SOUND AS AN EXAMPLE OF WAVE MOTION

a. How can we analyze the harmonic content of a sound wave?
b. What is an oscillograph?
c. What physical characteristics determine (1) the pitch, (2) the loudness, (3) the color of a sound?
d. What are supersonic waves?
e. How does sound intensity vary with distance from a point source of sound?
f. What is a decibel?
g. Is a sound which is 100 times as intense as another sound also 100 times as loud?
h. What is an echo?
i. How is the displacement at a point traversed by two waves determined by the individual wave motions?
j. What does the principle of superposition state?
k. What is meant by "interference" of waves?
l. Under what conditions do we obtain "standing waves"?
m. How do the condensations and rarefactions come about in longitudinal standing waves?
n. What relationship exists between wave length and length of a string along which standing waves are maintained?
o. To what frequencies will a pipe (open at both ends) resonate?
p. What are beats?
q. What is the Doppler effect?

Sound Vibrations

When we hold a book in our hands near a loud band, we can feel the book vibrate distinctly. The intensity of the vibrations which are detected by our fingertips varies with the intensity of the sound we hear. Sound is evidently capable of causing bodies to vibrate, and conversely, sound can be caused by vibrations of solids, liquids, and gases. Vibrating solids are very common sources of sound: gongs, drum membranes, strings, squeaky door bearings, new shoes, saw blades, human vocal cords (which can be considered as elastic membranes). Vibrating gas columns are used in organ pipes and in all wind instruments, and they play an important role in determining the sound of the vowels in speech. Vibrations originating in liquids are commonly heard in water-supply pipes when a faucet is turned on.

The generation and transmission of sound vibrations are of great importance as a means of communication used by diverse species of animals. Methods invented by man have extended the scope of transmission of sound messages to a practically unlimited range by the use of electricity. Figure 27.1 illustrates the principle of the **telephone** (*T*) and the **microphone** (M). The sound of the voice agitates the membrane D of the microphone M. The coil C is rigidly attached to D and thus participates in the vibration of the membrane. The magnet NS creates the magnetic field in which the coil C is vibrating. As a result of the fluctuations of the magnetic flux which penetrates the coil C, an alternating emf is induced in C. Thus, the velocity fluctuations of the vibrating membrane produce variations in emf whose pattern of frequencies corresponds exactly to the original sound. The varying emf induced in the coil of microphone M produces an alternating current in the coil C' of a distant telephone T. The telephone is shown to be identical in design with the microphone. The alternating current which traverses coil C' makes it behave like a bar magnet of varying strength which is alternately attracted and repelled by the stationary magnet SN with a periodically varying force. The frequency of fluctuation of this force is the same as the frequency of the alternating current flowing through C and hence is equal to the original sound frequency. Faithful reproduction of speech and music requires special adjusting circuits, since vibrations of the microphone coil C, which are of equal amplitude but different frequencies, will not give rise to the same amplitude of induced emf. According to the law of electromagnetic induction $V \propto \Delta\phi/\Delta t$. Hence the more rapid vibrations will give rise to a higher induced emf. This tends to distort the sound quality.

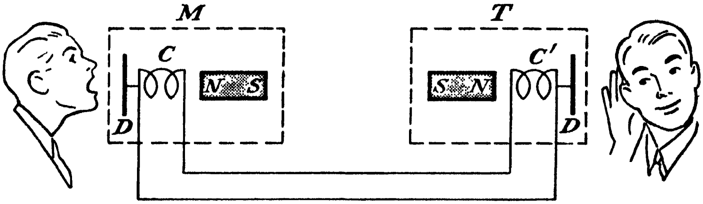


FIG. 27.1. Telephone and microphone.

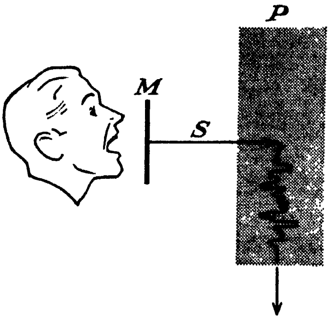


FIG. 27.2. Oscillographic recording of sound.

A membrane M could be attached to a stylus S resting on a glass plate P which is coated with lampblack. Figure 27.2 shows such an arrangement. As the plate moves downward, the vibrations are recorded continuously in a fashion similar to the performance of the kymograph discussed in Chap. 23. Such a device for the recording of rapid vibrations is called an **oscillograph.** Nowadays electrical oscillographs or oscilloscopes are used almost exclusively. The *cathode-ray oscilloscope* has been described in Chap. 23. The complicated curve recorded by the oscillograph represents the displacement of the vibrating membrane as a function of time. It can be resolved by Fourier analysis into component simple sinusoidal vibrations of different frequencies and amplitudes. Such an analysis enables one to determine the physical characteristics of sound responsible for what we refer to as the *pitch*, *intensity*, and *color* of sound (or timbre).

Sound Characteristics

It does not require an elaborate oscilloscope to show that the pitch of a sound is determined by the frequency of vibration of the source. For instance, if we brush a comb rapidly past the edge of a postal card, the card is set in vibration by the prongs of the comb and emits a sound of a definite pitch. If we increase the speed of the comb, the pitch of the emitted sound increases. We can determine the frequency which corresponds to a given pitch by a slight modification of this experiment. We replace the comb by a toothed wheel mounted on the shaft of a variable-speed motor. A sound will be produced when the edge of the card is held against the teeth on the rim of the rotating wheel. *The frequency of vibration of the card (and hence of the detected sound) is simply equal to the number of teeth which strike the card in a second.* If the frequency of rotation of the wheel is n per second and the number of teeth on the rim of the wheel is z, the frequency is given by $f = nz$. Doubling the angular speed of the wheel doubles the frequency f, and we hear a higher pitch, the so-called *octave* of the original tone.

The loudness or the intensity of the sound can be easily shown to be *determined by the amplitude of vibration.* An oscilloscope record reveals that an increase in loudness of the sound which sets the oscillograph membrane in vibration results in an increase of the amplitude of the latter.

But amplitude and frequency alone do not seem to be sufficient to describe a sound. We can distinguish, for instance, between the flute, violin, and the human voice even if these sources emit a sound of equal pitch and loudness. What is the distinguishing feature between these sounds?

An oscillograph record answers this question. Figure 27.3 shows the oscillogram of vibrations of a tuning fork, flute, horn, and the human voice producing sounds of equal pitch. We see that vibration of the tuning fork is very nearly sinusoidal. The vibrations of the other sources have the same period as the tuning fork; *i.e.*, the vibration pattern is repeated after the same time interval. Fourier analysis reveals that all these vibrations contain a sinusoidal component whose frequency is equal to the frequency of the tuning fork. The equality of this so-called **fundamental frequency** for these vibrations accounts for the fact that we ascribe the same pitch to them. Their difference lies in the fact that the harmonics are different in the case of the different instruments (different "overtones"). **Thus, the characteristic of a sound, the so-called *timbre*, or *tone color*, is determined by the frequencies and by the relative amplitudes (as compared with the fundamental vibration) of the overtones into which the complicated vibration can be resolved.** In the case of vowels the explanation of why a sung vowel "a" (as in father) sounds different from the vowel "o" (as in both) sung by the same person at the same pitch and why the vowel "a" is recognizable as such when sung at different pitch by different people is based on the fact that certain characteristic frequencies (the so-called *formants*), which bear no harmonic relationship to the fundamental frequency of the sung vowel, are found to be associated with the different vowel sounds.

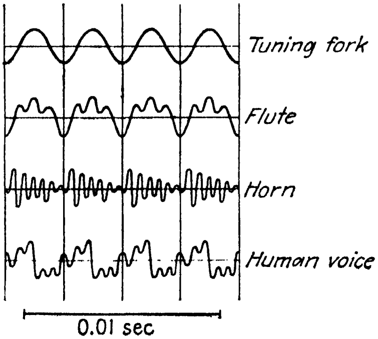


FIG. 27.3. Complex vibrations of different musical sound sources.

Supersonics

Experiments with sources of sound of variable frequency show that the range of sensitivity of the human ear extends from about 16 to about 20,000 cycles/sec, the upper limit being higher for young children. Certain animals are able to produce and hear sound of frequencies above 20,000 (**supersonic vibrations**). Dogs, for instance, can be called by supersonic whistles, which, because of the high frequency that they emit, are inaudible to human beings. This method is commonly used on the battlefield to signal to dogs without revealing one's presence to the enemy. Certain insects emit supersonic sound to communicate with each other. Phonographic recordings of supersonic mosquito sounds have been used to lure mosquitoes into a death trap.

Bats emit supersonic chirping sounds which enable them to avoid crashing into the walls of a dark cave in which they are able to fly about at great speed. The sound signals which are reflected from the walls and picked up by the bat's ears enable the bat to judge the location of the obstacle to be avoided.

The chemical and biological effects of intense supersonic vibrations show promise of important future applications. Thus, for instance, polymerized molecules like those of starch can be depolymerized by supersonic vibrations set up in a starch solution. Certain microorganisms, such as protozoa, can be killed, and some viruses, such as the tobacco mosaic virus, are rendered inactive. Heat developed in a medium through which supersonic vibrations are propagated suggests an alternative method of diathermy, *i.e.*, of heating internal organs.

Supersonic vibrations of 5×10^8 cycles/sec can be produced. Their wave length in air is 6×10^{-5} cm. This is of the order of magnitude of the wave length of light (see Chap. 29). This suggested the idea of a *sound microscope*, which may become useful in biological research. The object is "illuminated" by supersonic waves, and a "sound image" is produced by means of acoustical lenses or mirrors similar to those used in optical instruments (see Chap. 30). The acoustical image is translated into an electrical and then into a light image by means of a cathode-ray oscilloscope. Certain inhomogeneities which do not give rise to contrast with illumination by visible or invisible light (including X rays) can be detected with the help of the sound microscope.

Propagation of Sound

Sound requires a medium (such as air or water) through which to be propagated from the source to the ear. When we suspend a ringing bell in a container which can be evacuated, the bell can be heard only as long as the density of the air in the container is sufficiently high. When the container is thoroughly evacuated, the sound of the bell cannot be heard. The most powerful explosion on the surface of the moon would not be heard on earth!

Any gaseous, liquid, or solid medium is capable of transmitting sound. The particles of the medium are agitated by the source, and this agitation is propagated from particle to particle until the auditory mechanism of the ear is set in vibration. Thus, a conversation can be heard through a solid wall, and the sound which is produced by hitting two stones together under water can be heard above the water surface. The propagation of sound in water has found important applications in underwater telephony, in the detection of submarine obstacles, and in depth measurements, which are discussed on page 472. Conduction in solids is utilized in hearing aids which receive and amplify sound vibrations and are mounted behind the ear so as to transmit the vibrations to the inner ear by *sound conduction through solid bones.* The following simple experiment will demonstrate conduction of sound waves through bones: Take the handle of a teaspoon between your teeth, plug your ears with your fingers, and hit the other end of the spoon gently against a door. You will hear a loud sound reminiscent of church bells which has been transmitted through your teeth and bones to the inner ear.

Absorption of Sound

As sound travels through an extended solid or fluid medium, its intensity is diminished owing to two causes: (1) absorption, (2) dilution of the energy flow. To understand these phenomena we should consider an aspect of wave motion which has not been emphasized in the preceding chapter. The sphere A of Fig. 26.7 can be considered as our source. The energy of the vibrating sphere can be measured by its amplitude [see equation (25.9)]. As the wave issues from A, energy is imparted to the neighboring particles of our model of a medium. This energy must have been gained at the expense of the source A. To maintain the amplitude of A at a constant value we must supply energy to our source. *The propagation of the state of vibratory motion along our row of spheres can thus be considered as being associated with a flow of energy issuing from A.* We can adjust the energy supply to A so that each sphere along the infinitely long row of spheres receives as much energy per second from its neighbor on the left as it gives up to its neighbor on the right. Under these conditions the spheres can be maintained in a vibratory motion, the amplitude being the same along the entire row of spheres. But suppose that there was a frictional loss of energy in the transmission mechanism of our model. Then the amplitude of vibration would diminish from particle to particle as we recede from A. We express this state of affairs by saying that our medium **absorbs** the waves.

In the case of sound-transmitting media, we detect a more or less pronounced temperature elevation of the medium as a result of transmission of sound waves. This temperature elevation indicates "internal friction," *i.e.*, partial conversion of the energy of the orderly oscillatory motion of the particles of the medium into energy of disorderly heat motion. This "leakage" of energy, *i.e.*, *partial conversion of the energy of wave motion into heat, accounts for the marked absorption of sound in such media as sand, felt, and other so-called* **soundproof** *materials.*

The Inverse Square Law

The second cause of the diminution of sound intensity with distance from the source, which we termed "dilution of energy flow," can be understood as follows:

Imagine that a point source S (Fig. 27.4) represents a pulsating† sphere vibrating with a constant amplitude and thus imparting energy at a certain rate to the neighboring particles of the medium. The state of vibration will be propagated uniformly in all directions, and we can speak of an energy flow emanating from the source S. We can consider each spherical shell surrounding the source S through which this energy passes as a sort of window over the cross section of which the energy is distributed uniformly. *The density of energy flow, i.e.,* the amount of energy passing through each square centimeter of a spherical shell per second, is given by $E/4\pi R^2$, where E is the total amount of energy passing through the whole of the shell surface in a second. If there is no absorption, the same amount of energy will pass through a shell of radius R_2 as through one of radius R_1 each second: $E_1 = E_2 = E$. The density of energy flow, however, will be smaller at the distance R_2, since the energy flow is diluted here over a larger area. **Let us define intensity of sound I as the density of energy flow.** We can write then for the sound intensity at the distances R_1 and R_2

---

† That is, expanding and contracting.

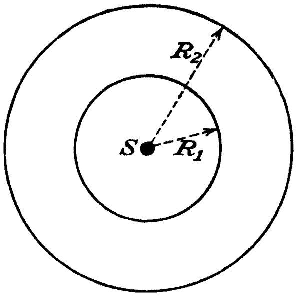


FIG. 27.4.

$$I_1 = \frac{E}{4\pi R_1^2} \quad \text{and} \quad I_2 = \frac{E}{4\pi R_2^2}$$

hence,

$$\boxed{\frac{I_1}{I_2} = \frac{E}{4\pi R_1^2}\,\frac{4\pi R_2^2}{E} = \frac{R_2^2}{R_1^2}} \tag{27.1}$$

If we refer to a fixed standard distance R_0 and put $I_0R_0^2 = C$, we can write

$$\frac{I}{I_0} = \frac{R_0^2}{R^2}$$

or

$$\boxed{I = \frac{I_0R_0^2}{R^2} = \frac{C}{R^2}} \quad \textbf{(Inverse square law)} \tag{27.1a}$$

The sound intensity diminishes in inverse proportion to the square of the distance from the source.

This law is an idealization; it does not ordinarily hold precisely under practical conditions because some of the assumptions underlying its derivation are not usually fulfilled. For instance, strictly nonabsorbing media do not exist. The usual sources of sound do not emit sound equally well in all directions. *The source of sound may be treated as a point source only at distances in comparison to which its size is negligibly small.*

Intensity and Loudness

The intensity of sound as defined above is measured in terms of energy per unit time per unit area, i.e., in watts per square centimeter. The range of sensitivity of the ear to acoustical energy flow is enormous. The faintest perceptible sound intensity of 1,000 cycles/sec corresponds to an energy flow of 10^{-16} watt/cm^2, and the loudest tolerable sound corresponds to an intensity of 10^{-4} watt/cm^2. The ratio of these limiting intensities is $r = 10^{-4}/10^{-16} = 10^{12}$. This does not mean, however, that the loudest tolerable sound appears to be one million million times as loud as the weakest audible sound. Experiments indicate that the sensation of loudness is not linearly proportional to the sound intensity but rather that it increases more nearly as the logarithm of the sound intensity. Thus, for instance, increasing the sound intensity a hundredfold only about doubles the loudness of the sound (log 100 = 2), and a thousandfold increase in intensity makes the sound about three times as loud (log 1,000 = 3)

This is an example of the psychophysical **Weber-Fechner law,** according to which *the intensity of a sensation is proportional to the logarithm of the intensity of the stimulus* rather than directly to the stimulus.

In practice, the loudness of a sound is compared with the loudness of an arbitrary reference sound of intensity $I_0 = 10^{-16}$ watt/cm^2. The logarithm of the intensity ratio of the two sounds which are being compared as to loudness is designated as **intensity level:**

$$\text{Intensity level} = \log_{10} \frac{I}{I_0} \text{ bells}$$

The intensity level is said to be 1 bell when $I = 10I_0$, so that

$$\log_{10} I/I_0 = 1$$

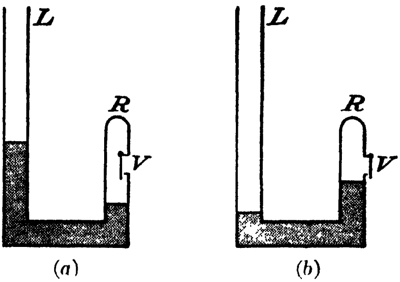
Practically, a unit ten times as small as a bell is used: the **decibel.** The above equation is then replaced by the following expression:

$$\text{Intensity level} = 10 \log_{10} \frac{I}{I_0} \text{ decibels}$$

Sound as a Longitudinal Wave Motion

Since transverse waves cannot be propagated in fluids, we assume that sound is propagated through air in the form of longitudinal waves in which, as we know, regions of progressing compression alternate with regions of rarefaction. This feature of longitudinal waves was elucidated in connection with the one-dimensional model of a medium (Fig. 26.9). In the three-dimensional case (Fig. 27.4) the longitudinal wave is to be pictured as a succession of regions of *condensation* (increased gas pressure) followed by regions of *rarefaction* (diminished gas pressure) which, like the regions of condensation, have the shape of spherical shells.

The existence of condensation and rarefaction in a progressing sound wave has been demonstrated by Quincke in an ingenious way. Imagine that the left leg of the water manometer of Fig. 27.5*a* is open and extends to a region free from sound waves. Assume that a sound wave passes the right (sealed) leg *R*, which is equipped with a valve *V*. To begin with, the water levels are equal in both manometer legs. The valve *V* is designed in such a way that the compressions which pass it push it open and allow the air to enter the leg from the outside. The rarefactions, on the other hand, close the valve, since the air in the process of escaping from *R* to the outside tends to close the valve, which remains closed as long as the inside pressure exceeds the outside pressure. The forcing of air into the leg *R* continues until the pressure in this leg has become equal to the maximum pressure in the regions of condensation. Thus, the manometer in Fig. 27.5*a* demonstrates compression in sound waves and allows the measurement of the value of the pressure amplitude.



(*a*) (*b*)

FIG. 27.5. Quincke manometers.

The valve *V* of the manometer in Fig. 27.5*b*, on the other hand, is designed so that it is closed when exposed to an external pressure and opened by an external rarefaction. The air continues escaping from *R* until the internal pressure has become equal to the pressure in the region of greatest rarefaction of the passing wave.

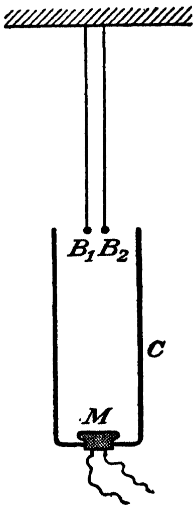


FIG. 27.6. Attraction of pith balls in a sound field.

The fronts of compression and rarefaction can be demonstrated very clearly by photography. A progressing sound wave is illuminated for a very short time interval by an electric spark. The succession of compressions and rarefactions gives rise to shadows on a photographic plate which reveals their configuration.

The to-and-fro motion of the air particles in a sound wave, as well as the longitudinal character of the wave, can be demonstrated by means of the following experiment (Fig. 27.6). B_1 and B_2 are very light pith balls suspended in pendulum fashion as shown in the figure. C is a glass cylinder, and M is a vibrating membrane. If we take out the membrane M, we can blow air upward through the opening. The pith balls are drawn then toward each other and collide. The same thing happens when we connect the lower opening to a pump and thus create a downward motion of air past the pith balls. We recognize in this demonstration the Bernoulli effect which was discussed in Chap. 12. The gap between the spheres acts like the constriction in a pipe. Air moves faster between the spheres than elsewhere, as a result of which the air pressure between the spheres is reduced and they collide. This effect is not influenced by a reversal of the direction of the wind.

Now when the membrane M is set in vibration, the spheres collide as before. We attribute this to a longitudinal up-and-down motion of air molecules which should have the same effect as a wind of alternating direction.

Reflection of Sound Waves

In Chap. 26 we have shown how a wave is reflected when it impinges upon a boundary between two media in which the velocity of propagation is different (Fig. 26.14). Since sound is a wave motion, we should expect to observe reflection of sound from boundaries between dissimilar media. Reflection of sound is commonly observed and is known as **echo.** Houses, groups of trees, hills, or boundaries between masses of air of different temperature may act as reflectors of sound.

Reflection of sound has been used to map the depth of the seas and oceans as follows: An acoustic signal (pulse wave) is sent out from the bottom of a ship. The wave travels to the bottom of the sea, is reflected, and returns to the bottom of the ship where the time of its arrival is recorded. The speed of sound in water being known, the depth is easily computed from the time required for the round trip of the pulse wave.

The same principle may be used to detect submarine obstacles or submarines. The primary pulse wave acts like an "acoustical searchlight." When it impinges upon a distant submarine, the latter reflects the wave impinging upon it and thus becomes a source of a reflected wave which is received by the search boat. This method may be used to determine the position and distance of the submarine. This is essentially the same principle that the bat uses to detect obstacles in a dark cave by listening to the echo of the sounds it emits. A method of detecting enemy airplanes and of determining the direction as well as distance of their location is based on the same principle except that, instead of acoustical waves, radio waves are used. The same procedure is also employed in geological prospecting. An explosive charge is detonated underground, and the subterranean echoes of the explosion are picked up at different listening posts. The time of arrival of these reflected sounds at the different stations enables one to determine the site of the reflectors, which are presumably rocks differing in composition from the surrounding ground.

Refraction of Sound Waves

Sound also exhibits the phenomenon of **refraction** like other waves. Let us consider first, for the sake of comparison, *refraction of ocean waves,* which explains why they usually arrive nearly parallel to the shore line. The velocity of waves in shallow water depends on the depth of the water. It increases as the depth increases. In Fig. 27.7 we see how the direction of the wave front gradually becomes parallel to the shore line owing to the fact that the lower portion of the wave front, which is closer to the shore line (and hence travels in shallower water), travels more slowly than the more distant portion. As a result, the wave front swings around like a row of soldiers in which the men farther away from the shore line walk faster than the men nearer the shore line.

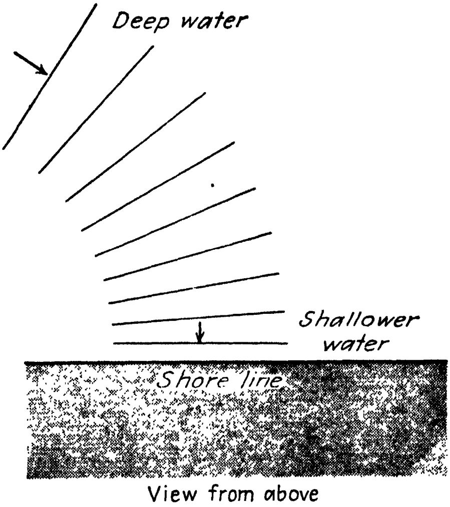


FIG. 27.7. Refraction of water waves.

A similar situation can be observed in the case of sound. At night the ground cools off rapidly by radiation and the air in the proximity of the ground is frequently cooled to temperatures lower than that of the overlying strata of air. Since it is known that the velocity of sound increases with the air temperature, we shall expect the velocity of sound to diminish as we proceed from the warmer top layers of air to the cooler ground layers. The conditions are similar to those in the case of water waves and shore line referred to in Fig. 27.7. In a sound wave whose front is originally vertical, the upper portions of the wave front will travel faster than the lower portions. The wave will be refracted groundward. This explains why sound is observed to travel farther along the ground at night than on a sunny day. During the daytime the conditions are reversed. The ground is rapidly heated by the sun and, in turn, heats the adjacent layers of air. Due to the existent gradual decrease of air temperature with height there is a gradual diminution in the velocity of sound propagation with height. The top of a vertical wave front travels more slowly than the bottom. As a result the sound wave is refracted upward and sound does not travel so far along the ground as it would have in a medium of uniform temperature.

Characteristic Feature of Wave Motion

A characteristic feature of wave motion is the transfer of energy without a simultaneous transport of matter. We shall see in subsequent chapters that it is not always easy to decide whether a transfer of energy is accomplished by wave motion or through a unidirectional motion of particles whose kinetic energy conveys power from a source to a receptor. Phenomena similar to reflection and refraction of waves can also be observed with swarms of particles and, hence, cannot be used as an indication of wave motion. An elastic sphere is reflected from a wall according to the same law of reflection which holds for a wave: *The angle of incidence is equal to the angle of reflection.* Moving particles, *e.g.*, electrons, can be deviated in electric fields and made to move along a curved path resembling the path of waves in a medium of varying index of refraction (see Fig. 27.7). It is very desirable to discover effects which could be considered *specific indications of wave motion* and which could be generally observed in cases of mechanical and nonmechanical wave motions. Such specific effects of waves have been discovered in the so-called phenomena of **interference of waves.** We shall study first interaction of sound waves and shall establish the pattern of behavior labelled as *interference.* We shall then assume in our subsequent studies that any phenomenon which follows the general pattern thus established is due to a wave motion.

Principle of Superposition

Imagine water waves emanating from a source to travel in the northerly direction toward an observer who measures the amplitude of the arriving waves. Now a second source begins to emit waves which travel eastward and pass through the region between the observer and the first source. How will this second train of waves influence the amplitude of the waves detected by the observer? Experiment shows that there is no influence! A similar experiment with sound yields the same result. **In general, the propagation of a wave is not affected by the passage of another wave through the same region!**

But now let us fix our attention on the region in which the two waves cross each other. What do we see? If we observe, for instance, the meeting of two waves, one being of a long wave length and the other of a short wave length, we have the impression of the short wave riding on top of the long wave. Such observations can be made with waves of any wave length meeting at any angle.

We could now perform the following quantitative experiment. We "turn on" one of our two sources of waves and observe at a selected point how far the water surface swings up and down; *i.e.*, we determine the amplitude of the transverse wave. A record of the excursion of the water surface as a function of *time* is represented by the sinusoidal graph *A* of Fig. 27.8. Now we "turn off" the first source and "turn on" the second source, which generates a wave of a different wave length and frequency. As a result the height of the water surface will vary as indicated by graph *B* of Fig. 27.8. How could we predict the instantaneous position of the water surface as a function of time if both waves were to pass our point of observation simultaneously? **Experience shows that the actually observed motion, which is represented by the solid curve *C*, is determined by adding the ordinates of the curves *A* and *B* algebraically at every point.** In other words, a point of our medium tends to perform simultaneously the motions which the individual waves tend to impose upon it (**principle of superposition**). In this fashion the agitation of the medium at a point can be determined with any number of waves passing that point.

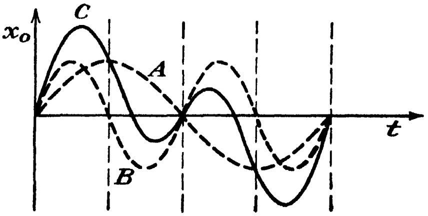


FIG. 27.8. Superposition of waves.

Interference of Waves

Let us now apply the principle of superposition to the important special case of two waves of equal amplitude and frequency, which penetrate each other, moving in opposite directions. We are confronted with such a situation, for instance, when a wave is reflected backward by a wall. In Fig. 27.9 the two waves meet at a point marked by the vertical solid line A. We can imagine them to be transverse waves moving along a liquid surface or along a stretched rope. We begin to take

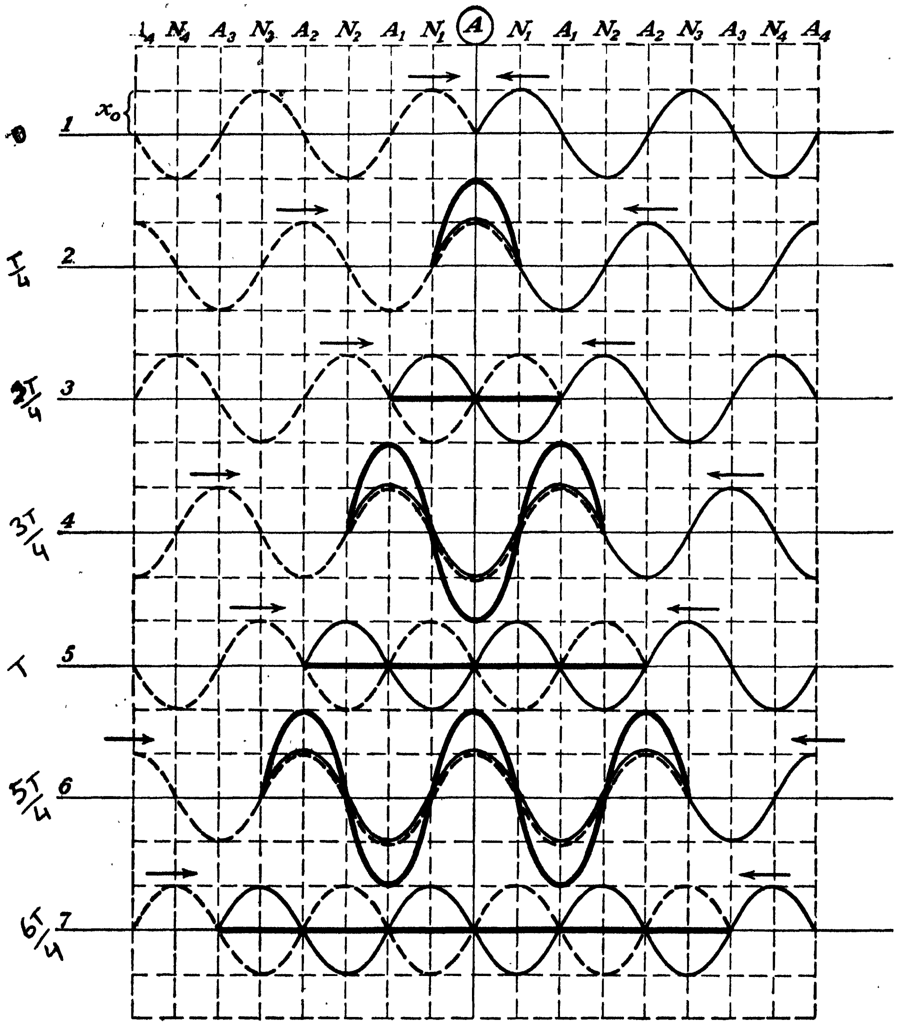


FIG. 27.9. Interference of waves.

snapshots of the region traversed by these waves at intervals of one-quarter of a period during which a wave advances by $\lambda/4$.

In snapshot 2 each wave is seen to have advanced a quarter of a wavelength in its direction of propagation. The region in which the two waves penetrate each other is situated between the first two vertical dotted lines marked by the letter N_1. According to the principle of superposition we obtain the excursion of the particles of the medium in this region of overlapping waves by algebraic addition of the ordinates of the coincident solid and dotted waves. The result of this addition is the deformation of the liquid surface (or of the string) indicated by the heavy curve. *The amplitude of this resultant wave obtained by superposition is equal to the sum of the amplitudes of the component waves.*

At the time of snapshot 3 the waves have advanced an additional distance of $\lambda/4$. The width of the region of overlapping is now seen to be equal to one wave length. At every point of this region the excursion which the particles of the medium would have had were the solid wave present alone is exactly equal but opposite to the excursion which would have been imposed by the dotted wave. As a result, the two waves cancel each other at all points between the two vertical dotted lines marked by letter A, and the liquid surface appears perfectly plane at this moment.

As we proceed to snapshot 4 and the following ones, we record the following sequence of events: In 4 the liquid surface shows again the maximum deformation. The region of interpenetration of the waves has been expanded by $\lambda/4$ in both directions. Whereas at the time of snapshot 2 the deviation of the particle located along the vertical solid line marked A was upward, at the time of snapshot 4 this particle exhibits a maximum downward deviation.

In snapshot 5 the surface is undisturbed again in the region where the two waves *interfere* with each other. And so as we proceed from snapshot to snapshot, we find pictures of maximum deformation of the surface alternating with photographs of an undisturbed horizontal level.

A closer inspection of Fig. 27.9 reveals that *not all parts of the liquid surface move up and down.* If we compare the successive positions of the liquid particles located at A by going down the vertical solid line, we see that these particles move up and down with the amplitude $2x_0$ if x_0 is the amplitude of the individual *progressing* waves. In snapshot 2 the liquid at A is at its highest point; it has fallen to the zero level in snapshot 3; in the next snapshot it has reached its lowest position $2x_0$ centimeters *below* the zero level; it is back to zero in snapshot 5 and has returned to the position of snapshot 2 in snapshot 6; *i.e.*, a full vibrational cycle has been completed.

But now let us examine the fate of a particle of liquid located at the intersection between the liquid surface and the line N_1. That particle is at rest at zero level in snapshot 2. It is at rest at zero level in snapshot 3. It is still in the same position in snapshot 4, and its state of motion remains unchanged in all of the following snapshots. **This**

point of the region traversed by both waves is called a *node*. The medium remains quiescent at all times at this point. It is the point at which the two wave motions always meet in phase opposition and hence persistently cancel each other. The quiescence at the nodes is in great contrast to the maximum agitation at all points marked by the intersection of the liquid surface with the lines A, A_1, A_2, A_3, etc. Let us consider, for instance, the position of the point at A_1 in the consecutive snapshots from 3 to 7. It starts from zero, moves up, down to zero, down below zero, and back to zero in 7. **The regions of maximum agitation are referred to as *antinodes*.**

Transverse Standing Waves

How do the intermediate points move which are located between the nodes (marked by N, N_1, N_2, etc.) and antinodes (marked by A, A_1, A_2, etc.)? They are located between points N, which can be said to vibrate with the amplitude $x_N = 0$, and points A, which move up and down with amplitude $x_A = 2x_0$. The maximum excursion of the intermediate particles increases from 0 to $2x_0$ as we proceed from points N to points A. The resulting pattern of vibration can be easily observed by watching a vibrating string or rope (Fig. 27.10). A vibrator S moves the left end of the rope, which is tied to a wall at W, up and down.

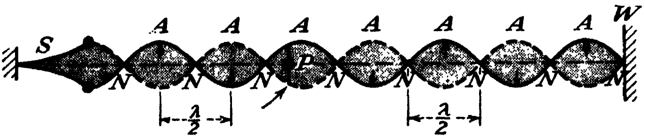


FIG. 27.10. Stationary transverse rope waves.

A wave moves along the rope toward W, and a reflected wave travels from W back toward the source S. The interference of these waves produces a pattern of nodes and antinodes. The point tied to the wall is necessarily a node.

Due to the rapid vibratory motion, we have the impression of stationary loops, which are delineated by the dashed and the solid intersecting sinusoidal lines in Fig. 27.10. The vertical distance between these two lines determines the amplitude of the string at a particular point (see, for example, particle P marked by arrow in Fig. 27.10). Figure 27.9 reveals that the vibrations in two adjacent loops are in phase opposition: Whereas the particle at A_1 in snapshot 6 moves down, the center particles at A and A_2 in the neighboring loops move upward. It can be also inferred from Fig. 27.9 that **the distance between two adjacent nodes or antinodes is equal to half the wave length.**

This vibrational pattern has a wavelike appearance, but the wave does not appear to travel along the x axis as do the two component waves which create it by interference. The wavelike pattern appears to stand still and is for this reason called a **standing wave.**

Such standing waves are of great importance in various branches of physics. **We consider our ability to produce a standing-wave pattern as convincing evidence proving the wave nature of a phenomenon.**

Longitudinal Standing Waves

The results obtained in the study of standing transverse waves can be transferred to the case of longitudinal waves. Owing to the difficulty of visualization and graphical representation of longitudinal waves the same diagram is used in connection with them as in connection with transverse waves. The *longitudinal motion of the particles can be inferred from a transverse diagram according to the following scheme of "translation":* Fig. 27.11 shows two snapshots of a string which exhibits a stationary **transverse-wave pattern.** At the time zero, the left half of the string is bent upward, whereas the right half is depressed (Fig. I). Half a period later the right half is elevated, whereas the left half is

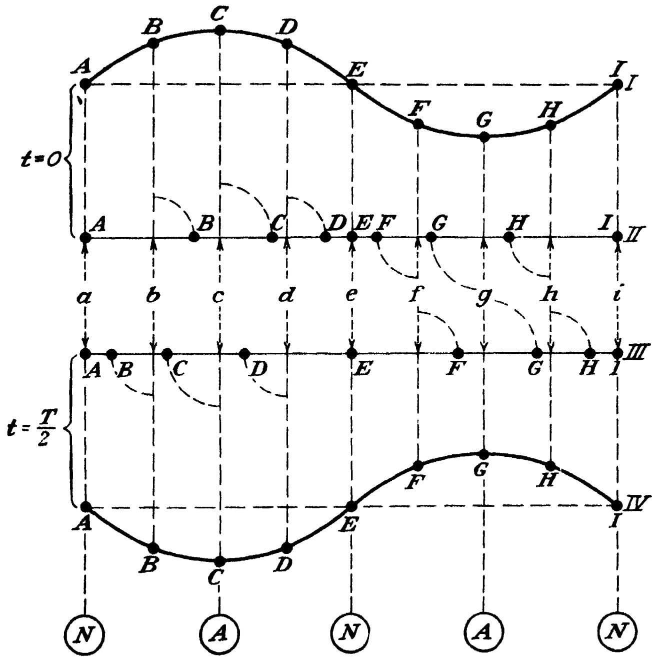


FIG. 27.11. Transverse representation of longitudinal wave motion. The equilibrium positions a, b, c, . . . are indicated by arrows.

lowered (Fig. IV). It is easy to visualize how the shape of the string is being distorted throughout the complete cycle as the individual particles vibrate up and down, the nodes and antinodes of displacement being marked by N and A, respectively.

In a longitudinal vibration the individual particles (some representative particles of the string have been marked by $A, B, C, \ldots, I$) move to and fro in the direction of the string just as the particles in the illustrated transverse motion move up and down. *We can visualize how the particles in a longitudinal standing wave would vibrate by the following method of plotting their positions:* Assume that the *upward* deviation of a particle in the transverse wave corresponds to a deviation toward the *right* in a longitudinal wave, a *downward* transverse displacement to a longitudinal displacement to the *left*. Then the instantaneous positions of the particles in a *longitudinal* wave can be obtained from their illustrated positions in the *transverse* wave (curve I of Fig. 27.11) by drawing a circular arc (in Fig. II) about the points $a, b, c, \ldots$ (equilibrium positions of the particles) with a radius which is equal to the transverse deviation of the particle. Thus, the longitudinal displacements of particles $A, B, C, \ldots, I$ are obtained (line II of Fig. 27.11) by converting the upward deviations of wave I into deviations to the right along line II and the downward deviations of wave I into left deviations along the line II.

We see then that at $t = 0$ (moment of maximum transverse distortion in Fig. I with a node at E) the longitudinal standing wave (line II) exhibits a region of maximum condensation at the node E, at the point where the particles do not move in the longitudinal wave any more than they do in the transverse wave. This should be the region of maximum pressure in a sound wave. (The original undisturbed positions of the particles are marked by small letters $a, b, c, d, \ldots$.) At the other two nodes (undisplaced particles at A and I) we notice regions of rarefaction (minimum pressure) characterized by a greater separation of neighboring particles than in any other region (note the distance between A and B as compared with distance ab). Thus, we encounter alternating regions of condensation and rarefaction as we proceed from node to node.

In order to see where the particles move to from their positions occupied at the time $t = 0$, we examine the instantaneous picture taken half a period later. The transverse displacements shown in IV are again translated into longitudinal displacements according to the scheme outlined above. The position of the particles at this moment is shown on line III of Fig. 27.11. We see that at the node E, where a condensation was recorded at the time $t = 0$, a rarefaction now exists, whereas the nodes at A and I now are points of condensation (note the distance between A and B as compared with ab).

We infer from this that the particles must have moved as follows during this half-period interval: all particles on the left of E away from E toward A and all the particles to the right of E away from E toward I. This leads to a diminution of density at E and to its increase at A and I.

The consideration of subsequent snapshots (at the times $t = T$, $t = \frac{3}{2}T$, $t = 2T$, etc.) reveals that *the vibrating particles move periodically toward a node and away from it. The closer to a node a particle is located the smaller are its excursions.* The greatest excursions are observed in the regions of the antinodes, but in these regions the relative distance of the particles, *i.e.*, their density, practically does not change. Thus **whereas the nodes are regions of maximum pressure variation and absence of vibration, the antinodes are regions of constant pressure but of maximum agitation of particles.**

Stationary-wave Patterns

Let us now confine our attention to the pattern of a standing wave. Figure 27.12 shows a typical example of a standing wave in a vibrating string. The two ends at which the string is attached to the walls W_L and W_R are necessarily nodes of displacement. In analogy to Huygens' principle we can consider, for instance, the point P of the string as a source from which waves of the frequency f are moving in both directions. They are reflected when they reach the walls and return moving in opposite directions. Since the point of the string attached to a wall is motionless, *we conclude that the wave arriving from P at the point of attachment and the reflected wave must be in phase opposition* at this point, so that the string remains here at rest owing to the destructive interference between the incoming and the reflected wave: **The reflection of the wave from the wall is accompanied by a phase reversal.**†

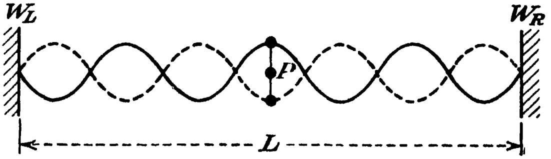


FIG. 27.12. A stationary-wave pattern.

With the two reflected waves emerging from W_L and W_R and moving in opposite directions, we have the conditions which lead to the generation of stationary waves as illustrated in Fig. 27.9. The region of overlapping of these waves is the length L of the string.

---

† A similar result is obtained if the string is attached to a heavier rope instead of to a wall.

Will a standing-wave pattern be established for *any* frequency of vibration of the point P? The answer is evidently no. Since the ends of the string must be nodes, the inspection of Fig. 27.12 shows that this will happen *only at frequencies at which the string is divided into an integral number of loops, so that the length L is an integral multiple of* $\lambda/2$ where the wave length λ is determined by the frequency f and the velocity of wave propagation v as follows: $\lambda = v/f$.

Hence,

$$\boxed{L = n\frac{\lambda}{2} = n\frac{v}{2f}} \tag{27.2}$$

The frequencies at which standing waves will persist are therefore given by

$$\boxed{f = n\frac{v}{2L}} \tag{27.3}$$

The observed frequencies form a series of integral multiples (n being an integer) of a lowest frequency (**harmonics of the fundamental frequency**). This lowest frequency f' is observed in the case where the minimum possible number of nodes (two at the walls) are encountered along the string. In this case $L = \lambda/2$, and hence, $n = 1$. Thus we get $f' = v/2L$.

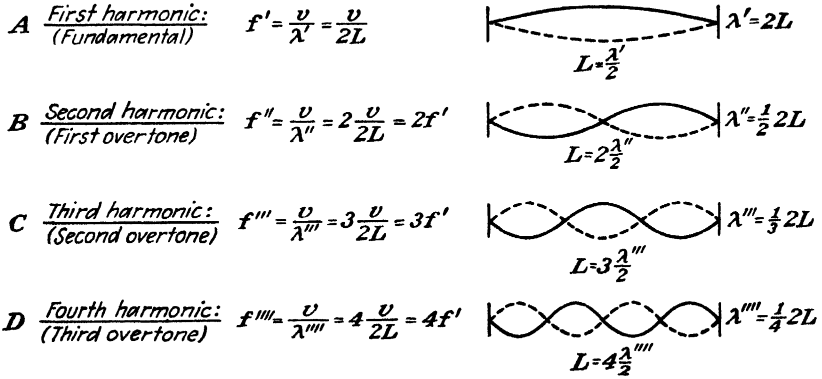

A
First harmonic: (Fundamental)
$f' = \frac{v}{\lambda'} = \frac{v}{2L}$
$\lambda' = 2L$
$L = \frac{\lambda'}{2}$
B
Second harmonic: (First overtone)
$f'' = \frac{v}{\lambda''} = 2\frac{v}{2L} = 2f'$
$\lambda'' = \frac{1}{2}2L$
$L = 2\frac{\lambda''}{2}$
C
Third harmonic: (Second overtone)
$f''' = \frac{v}{\lambda'''} = 3\frac{v}{2L} = 3f'$
$\lambda''' = \frac{1}{3}2L$
$L = 3\frac{\lambda'''}{2}$
D
Fourth harmonic: (Third overtone)
$f'''' = \frac{v}{\lambda''''} = 4\frac{v}{2L} = 4f'$
$\lambda'''' = \frac{1}{4}2L$
$L = 4\frac{\lambda''''}{2}$

Fig. 27.13. Modes of vibration of a string.

Figure 27.13 is a survey of the frequencies which correspond to the different modes of vibration of a string.

We see that an infinite number of harmonics is possible, their frequencies being integral multiples of the fundamental frequency.

The ability of a string to vibrate in different modes with a discrete sequence of frequencies distinguishes this type of oscillator from a simple harmonic oscillator such as a pendulum. A pendulum has only one natural frequency with which it vibrates when disturbed. But a string has an infinite number of natural frequencies, namely, all the harmonics. When we distort the string as shown by the dotted line in Fig. 27.13*a* and release it, it will vibrate with the fundamental frequency. If we distort the string as shown by the dotted line in Fig. 27.13*c*, it will vibrate with the frequency of the third harmonic, $f''' = 3f'$, etc.

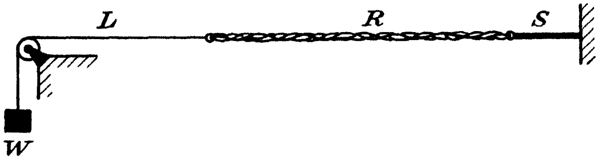


FIG. 27.14.

Accordingly, *the string will also resonate to this multitude of natural frequencies,* so that if we sound a tone of a frequency corresponding, for instance, to the fourth harmonic, the string will vibrate in the pattern shown in Fig. 27.13*d*.

A heavy rope *R* attached to a light string *L* (Fig. 27.14) and to a vibrator *S* on the other side will show a different pattern of vibration (*L* represents a medium of a smaller linear density than *R*). The relatively free end where *R* joins *L* is not a node but rather an antinode of displacement. This implies that the **waves which are propagated along the rope *R* are reflected at the less dense medium *L* without a phase reversal.**

It is left to the student as an exercise to determine what harmonics are emitted by a transversely vibrating thin rod clamped *at one end* when it vibrates in various modes as illustrated in Fig. 27.15.

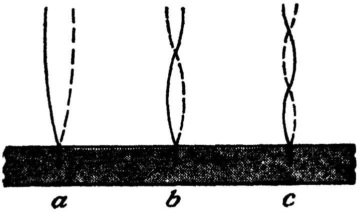
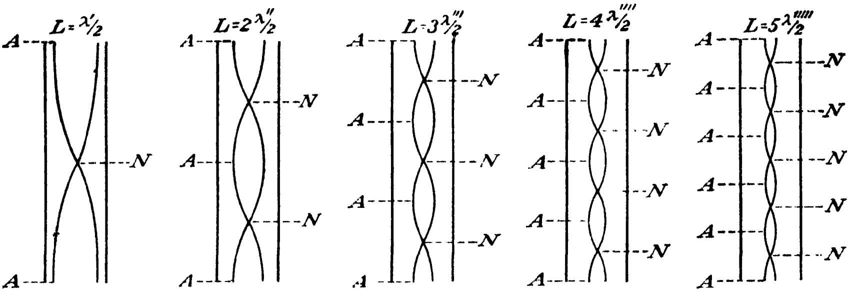


FIG. 27.15. Modes of vibration of a clamped rod.

Air columns enclosed in pipes vibrate in a fashion similar to strings and rods. If the air is enclosed in a cylinder which is *closed at both ends,* the two ends form nodes of displacement, since the covers of the tube prevent the air particles from moving back and forth longitudinally. As we have seen before, however, nodes of displacement are points of maximum pressure changes! Hence, the pressure variations will be at a maximum at the closed end of a cylindrical gas column. If, on the other hand, the cylinder is *open at one or at both ends,* the open end will be in communication with the atmospheric air and the pressure will remain constant there whereas the air particles will be free to move back and forth, so that the open end will be an antinode of displacement.† Figure 27.16 illustrates the first five stationary-wave patterns which can occur in a cylinder open at both ends.



(a) (b) (c) (d) (e)

FIG. 27.16. Modes of vibration of the air column in an open cylinder. (The antinodes are antinodes of displacement.)

Resonance

Air columns can resonate to different harmonics just as strings do. Resonance of air columns forms the basis of the wind instruments and also plays an important role in speech. The character of our voice is determined to a great extent by resonance of the gas volume in the cavities of our mouth, throat, and nose. This can be demonstrated by inhaling hydrogen (or, more safely, helium) and by speaking while exhaling the gas. The voice is transformed beyond recognition owing to the fact that the velocity of sound in these gases is much higher than in air so that the resonance frequency of the cavities is altered by changing the gas in it:

The connection between a resonance frequency of a gas column‡ and the velocity of propagation of sound in it is the same as the relation between a harmonic frequency of a string and the wave velocity in it as expressed by equation (27.3).

Our speech depends upon our ability to change the resonant frequency of the cavities of our speech organs. The distinction among the vowels is achieved by changes in size and shape of our mouth cavity through the motion of lips, tongue, and jaw. When a frequency to which the mouth cavity is adjusted to resonate is present in the sound emitted by the vocal cords, this frequency is amplified above the other Fourier components of the emitted sound. The different vowels are produced by varying the resonance frequencies of our mouth cavity.

---

† Theory and experiment show that the antinode is not located exactly at the end of the cylinder but that it lies roughly $R/2$ centimeters outside the tube (where R is the cylinder radius).

‡ An "open pipe" as shown in Fig. 27.16, a "closed pipe" in Fig. 27.18.

The excitation of vibrations in an air column by resonance can be demonstrated by the following simple experiment (Fig. 27.17).

A tuning fork T, emitting a sound of a definite frequency, is held over the opening of a glass tube. The tube can be filled to the top with water by raising the reservoir R. As the liquid level in the tube G is lowered by lowering R, the water level reaches a position P_1 in which we hear a louder sound. As we pass this point, the sound gets weaker but then increases in intensity sharply when a new position P_2 is reached by the liquid level. As we move the liquid level lower, the sound weakens again but increases sharply when the water level in G reaches the mark P_3, etc.

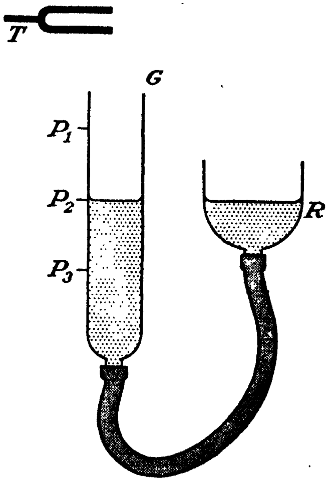


FIG. 27.17. Demonstration of resonance of an air column.

The explanation of this effect of intensification is suggested by the following observation: We note that the distances P_1P_2 and P_2P_3 are equal to $v/2f$ (where v is the velocity of sound in air and f the frequency of the tuning fork) and the distance from P_1 to the top of the tube is nearly equal to $v/4f$. Since $v = f\lambda$, we see that $P_1P_2 = P_2P_3 = \lambda/2$ and the distance from P_1 to the top is $\lambda/4$. We are evidently observing here

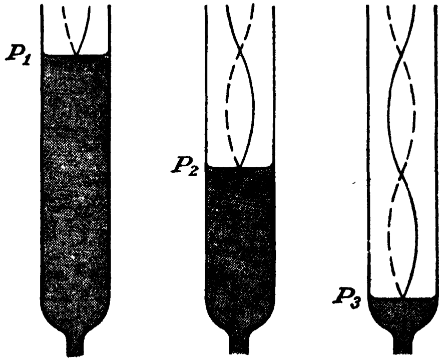


FIG. 27.18. Modes of vibration of the air column in a cylinder open at one end. (The nodes are nodes of displacement.)

resonance of the air column in G vibrating in different modes shown in Fig. 27.18 in transverse representation. Since the air column is closed at the bottom, the liquid level is always a node whereas the open end is always an antinode of displacement.

At resonance the air column drains energy from the tuning fork at a rapid rate and reradiates it into space. As a result energy escapes from the tuning fork and the vibrating air column at a more rapid rate than from the vibrating fork alone, and hence the intensity of the emitted sound is greater.

Beats

In the preceding considerations we dealt with the interference of waves of equal frequency. **The case where the two interfering waves are of different frequencies** is also of great practical importance. Let us imagine two waves of frequencies f_1 and f_2 crossing each other at a certain point P. According to the principle of superposition, the particles of the medium at point P will have an excursion equal to the sum of the individual excursions which would have been produced by each of these two waves in the absence of the other. In other words, the particles perform two vibratory motions simultaneously which are imposed upon them by the two interfering waves of different frequencies (f_1 and f_2).

We can form a clear picture of how a particle of the medium will move under such conditions by considering the following example: Imagine two pendula of *slightly* different frequencies f_1 and f_2. Assume that one pendulum is mounted in front of the other and that both of them are displaced by the same angular amount and then released to swing freely. At first the two pendula will seem to swing in phase, but then a moment will come when pendulum 1 swings to the right while pendulum 2 swings to the left. For example, if the frequencies are $f_1 = 10$ swings per minute and $f_2 = 9$ swings per minute, the pendula will start out in step and will again be in step 1 min later, after the first one has completed 10 swings and the second one 9 swings. In the middle of this minute interval, however, the two pendula will be roughly in phase opposition swinging in opposite directions.

If $f_1 = 110$ swings per minute and $f_2 = 100$ swings per minute, it is easy to see that the pendula will be 10 times ($f_1 - f_2$ times) in phase during this minute, namely, for the first time after the eleventh swing of pendulum 1 and the tenth of pendulum 2, then again at the twenty-second swing of pendulum 1 and the twentieth of pendulum 2, etc., and they will be just as often exactly out of step.

If one single pendulum were to perform simultaneously the motion of both these pendula, how would its motion look? At the moment when the pendula are *in phase,* the two equal amplitudes add up to a double amplitude, but at the moment when the pendula swing *in phase opposition,* the resultant amplitude is zero. If two wave motions were to affect a particle of the medium simultaneously, they would tend to make it move like the above-mentioned pendulum. The particle would vibrate with a variable amplitude, which would vary from $2x_0$ (where x_0 is the amplitude due to either single wave separately) to zero at the moment when the waves are in phase opposition. The frequency of this variation of amplitude would be equal to the frequency difference, $f_1 - f_2$, of the two waves. *If the interfering waves are sound waves, a tone of rapidly fluctuating loudness should be heard;* this is actually the case. *These fluctuations of loudness are called* **beats.** They can be demonstrated simply when two people whistle simultaneously two notes of slightly different pitch. **The beat frequency is equal to the difference between the frequencies of the interfering waves.**

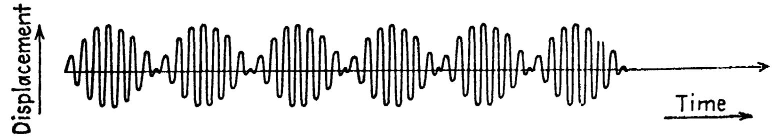


FIG. 27.19. Vibration pattern due to interference of two waves of unequal frequency.

Figure 27.19 shows the displacement as a function of time at a point where two waves of different frequencies interfere, giving rise to beats: When the two frequencies are sufficiently far apart, the difference frequency $f_1 - f_2$ may be high enough to give the sensation of a sound. The violinist Tartini discovered when playing simultaneously on two strings ("double stops") that a third tone could be heard.

Beats are utilized in the so-called *heterodyne* radio circuit in order to produce electromagnetic oscillations of relatively low frequency by allowing the high-frequency radio wave received from the transmitter to interfere with a wave of slightly different frequency generated in the receiver, thus giving rise to beats of a desired frequency.

Doppler Effect

When we listen to a sound produced by a source which vibrates with a definite frequency, we hear the same sound whether we are located in an atmosphere of hydrogen or nitrogen, under water, or behind a solid wall. As the sound reaches our ear traveling through different media, its wave length is changed according to the relation $v = f\lambda$ but its frequency remains unaltered. Does it mean that we must hear the same frequency under all circumstances as long as the frequency of the source remains unaltered? This is not the case. When we hear a sound of the frequency f, we are merely sure that f condensations (pressure maxima) pass the opening of our ear in a second. It is obvious that we could change this figure without affecting the frequency of the source. We could, for instance, move away from the source with the speed of sound, so that the same condensation would always remain in front of our ear opening, or we could move at a slightly slower rate, so that the sound waves would overtake our ear very slowly and would enter it at the rate of a few condensations per second.

In other words, *motion of the observer relative to the source of sound could affect the observed frequency.* It is evident that *such an effect should be observable with a moving source as well as with a moving observer.* Anyone who listens carefully to the sound of an approaching insect will notice a sharp drop in the pitch of the sound as the insect passes his ear. The change in pitch due to the relative motion between the source and the observer is called the **Doppler effect** after its discoverer.

We can perform the following experiments to demonstrate the Doppler effect:

Experiment 1. We station two cars along a highway and let car 1 blow its horn continuously while both car 1 and car 2 are at rest. As soon as the observer in car 2 starts moving toward car 1, he will hear a sound of *increasing pitch* as his car *accelerates* toward car 1. As soon as car 2 has reached a *constant speed,* the pitch as heard by the observer in car 2 will remain *constant* but *higher* than the pitch heard when car 2 was at rest. As car 2 approaches car 1 at constant speed, the sound gets louder but there is no change in pitch.

Experiment 2. We make the observer *drive away* from the source in car 1. He will then hear a sound of a *lower pitch* than the one he heard while at rest. Again the pitch heard depends only on his speed but not on his distance from the source.

Experiment 3. Now the observer in car 2 remains at rest while car 1 drives toward him. The effect is *qualitatively* the same as in experiment 1.

Experiment 4. The observer is still at rest, but car 1 moves away from him. The effect is *qualitatively* the same as in experiment 2.

We shall now treat this phenomenon *quantitatively* in order to derive an expression for the change in the observed pitch.

Assume in the following consideration that the reference system is at rest with respect to the medium (air).

Case 1. **The source is moving, the observer is at rest.**

Assume the observer B to remain at rest with respect to air which, in turn, is at rest with respect to ground. A source initially in position S_1 (Fig. 27.20) emits continually a sound of frequency f. Figure 27.20

shows a snapshot of this situation taken after the source has been moving for 1 sec with velocity S centimeters per second toward the observer B. The final position of the source is S_2, and the distance S is equal to the speed of the source. The initial distance between the source at S_1 and the observer B has been chosen for simplicity of derivation so that the first wave front (marked by 1 in the figure), which was emitted by the source while at S_1, reaches B exactly 1 sec later, namely, at the moment when this snapshot is taken. The last of the f wave fronts which are emitted by the source in 1 sec, is emitted at this moment by the source which is now at S_2. The f wave fronts emitted during this second are distributed uniformly between S_2 and B over a distance of $(V - S)$ centimeters, where V is the distance the sound travels in 1 sec.

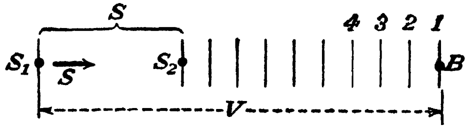


FIG. 27.20.

The wave will now begin to pass the ear of the observer, which is located at B. How long will it take these f waves to pass B? They travel with the speed V, and the distance the last wave must cover to reach B is $d = V - S$; hence, the time for passing B is

$$t = \frac{d}{V} = \frac{V - S}{V} \qquad (27.4)$$

The apparent pitch p heard by the observer B is given by the number of vibrations passing his ear in a second. We just determined that f vibrations will pass his ear in $t = (V - S)/V$ seconds. Hence, the frequency he hears is given by

$$p = \frac{f}{t} = \frac{f}{(V - S)/V} = f\frac{V}{V - S} \qquad (27.5)$$

In case the source recedes from B its velocity changes sign and we have to use $-S$ instead of S in the above derivation. We obtain the same expression for p as above except for a positive sign in front of S. We can write a general expression applicable to both of these cases as follows:

$$\boxed{p = f\frac{V}{V \pm S}} \qquad (27.6)$$

where V is velocity of sound; S is velocity of source; p is frequency of sound heard by stationary observer; f is frequency of sound emitted by moving source. The negative sign is chosen when S approaches B, the position sign when S recedes from B. This means **the frequency is elevated in the case of approach and lowered in the case of recession.**

Case 2. **The observer is moving, the source is at rest.**

We consider a snapshot taken after the observer B has been moving for 1 sec toward the source S with the velocity b (Fig. 27.21). Then the distance between his initial and final position at the end of the second is $B_1 - B_2 = b$. If the observer had remained at rest at B_1, he would have been passed continually by f waves each second. But owing to his motion toward B_2, he will pass an additional number of waves during this second. We can determine this additional number of waves by imagining the waves, which have been emitted from S, to stand still for a second while B moves past them from B_1 to B_2. How many waves did B pass on his way? There are f waves emitted by S in a second, which are spread over a distance of V centimeters; hence, there are f/V waves per centimeter. B has covered the distance of $B_1 - B_2 = b$ centimeters; hence, he has passed $n = b(f/V)$ waves. Thus the frequency p which he hears is given by

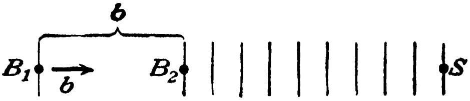


FIG. 27.21.

$$p = f + b\frac{f}{V} = f\left(1 + \frac{b}{V}\right) = f\left(\frac{V}{V} + \frac{b}{V}\right) = f\left(\frac{V + b}{V}\right)$$

If B recedes from S, the same expression is obtained except for a negative sign in front of b (velocity of observer). We can write, in general,

$$\boxed{p = f\left(\frac{V \pm b}{V}\right)} \qquad (27.7)$$

The positive sign is chosen in the case of approach, and the negative sign in the case of recession.

Doppler effect can be observed with all types of wave motion. We shall discuss some very interesting astronomical applications of the Doppler effect in the section on light (Chap. 31).

APPENDIX I

The Musical Scales

A variety of sequences of sounds, known as *musical scales*, has been used by various peoples, such as the Arabian scale, the pentatonic scale (used by the Chinese), and the diatonic scale used in occidental music.

As an example of a scale let us consider the diatonic C major scale:

Name of the note	C	D	E	F	G	A	B	C′
Frequency	264	297	330	352	396	440	495	528
Frequency ratio		9:8	10:9	16:15	9:8	10:9	9:8	16:15

In this tabulation the frequency of vibration is recorded under the letter symbol of the musical note and the ratio of two consecutive frequencies in the line below. These ratios are called *intervals.* The ratios 9:8 and 10:9 are the full-tone intervals and the ratio 16:15 corresponds to the half-tone interval.

The following intervals play an important role in musical harmony:

Intervals	Notes	Frequency ratio
Perfectly consonant:		
Octave	C′ C	2:1
Fifth	G C	3:2
Fourth	F C	4:3
Imperfectly consonant:		
Major third	E C	5:4
Minor third	G E	6:5
Major sixth	A C	5:3
Minor sixth	C′ E	8:5
Dissonant:		
Second	D C	9:8
Major seventh	B C	15:8
Minor seventh	C′ D	16:9

We see from this tabulation that the degree of consonance of an interval increases as the integral numbers expressing the ratio decrease. The most perfectly consonant interval, the octave, has a frequency ratio expressed by the two smallest integers.

In order to enable one to reproduce all the possible diatonic scales with a good degree of approximation without the use of an excessive number of different notes, the keyboard instruments utilize an octave divided into 12 equal intervals. There are 12 half-tone intervals in an octave such that the frequency ratio between two consecutive notes is $\sqrt[12]{2}$. If we start with the ground tone of the scale and multiply its frequency by $\sqrt[12]{2} = 1.0595$, we obtain the frequency of the note half a tone above the ground tone. If we repeat this procedure 12 times, we obtain the frequency of the higher octave.

This sequence of tones is referred to as the *even-tempered scale.*

APPENDIX II

Additional Examples and Applications

1. The student is referred to textbooks on physiology for study of the **anatomy and mode of functioning of the human ear.** We shall merely highlight here a few points of physical interest.

In the ear, conduction of sound in gases, liquids, and solids is utilized. The gas is the air filling the passages of the outer ear and the middle ear. The solids are the bones of the middle ear ("hammer," "anvil," and "stirrup") which transmit the vibrations from the eardrum to the oval window of the cochlea. The cochlea is filled with a liquid which is agitated by the vibrations of the oval window. The cochlea is a snaillike spiral shell which is divided into two parallel spiral half shells by a partition, part of which is the basilar membrane. This membrane resembles a small carpet lined with about 23,000 tiny "rods of Corti." These rods are agitated by the vibratory motion of the surrounding liquid. The agitation of each rod stimulates a particular nerve ending communicating with it. Each rod is sensitive to a different frequency of sound. The rods nearest the oval window respond to the highest frequencies. The frequency of response diminishes toward the distal end of the basilar membrane.

2. Our ability to determine the direction from which sound comes is due to the fact that we have two ears. Sound coming from the right reaches the right ear first. The magnitude of the very tiny time interval between the arrival of the signal at the two ears allows us to judge rather accurately the direction of the source. **Directional hearing** depends on the difference of sound intensity and on the phase difference of the vibrations at the two ears.

3. If the oscillograph of Fig. 27.2 is modified by making the stylus scratch its curve into a wax plate rather than write on a smoked glass plate, we obtain a so-called **phonograph.** If the stylus is replaced in the groove which it scratched into the wax and the plate moved down again, the membrane will be made to reproduce the vibration which originally gave rise to the curved groove. It will therefore emit the same sound as was used to operate the phonograph membrane. This is the principle of the **gramophone.**

The uses of phonographs and gramophones are well known. Recently, applications have been made in medical teaching. Important diagnostic sounds (such as heart sounds) of a large number of typical cases can be demonstrated by means of acoustical records to a large audience. Previously, accidental arrival of a typical case after a long waiting time and individual examination of the patient were the only available means of clinical demonstration.

4. Propagation of sound waves through the tissues of the human body is utilized in various diagnostic procedures. Sound produced in the larynx (by repeating certain words) is propagated down the trachea and reaches the lungs via the bronchi. The sound vibrations thus propagated travel through the tissues separating the lungs from the surface of the body and can be felt by one's fingertips held against the chest (**palpation**). If, for instance, one of the lobes of the lung is consolidated, sound is transmitted through it more readily than through the healthy lobe, which is filled with air. As a result, the vibrations detected by palpation over the diseased area are stronger than those near the normal portion of the lung. This symptom is used in the diagnosis of lobar pneumonia.

Palpation is also of value in the diagnosis of heart disease. For instance, in the case of mitral stenosis the passage of the blood through the constricted mitral valve of the heart into its left ventricle gives rise to a presystolic thrill, which evokes a sensation in the palpating hand similar to that felt when a purring cat is being stroked.

5. Listening by direct ear contact with the chest wall or through a **stethoscope** (a tube with a bell-shaped end which is applied to the patient's chest) is a very important diagnostic procedure (**auscultation**). The sounds produced in the lungs during the process of breathing can lead to the diagnosis of various pulmonary disorders.

Hippocrates was perhaps the first to have used this diagnostic method. He describes, for instance, how it is possible to detect fluid in a patient's chest. By pressing the ear against a patient's chest and shaking him one can hear a splashing sound.

The observation of **heart sounds** is also of great importance, since they enable one to diagnose various types of heart disease. Harvey was one of the first to have called attention to the heart sounds. He writes: "It is easy to see when a horse drinks that water is drawn in and passed to the stomach with each gulp, the movement making a sound, and the pulsation may be heard and felt. So it is with each movement of the heart, when a portion of the blood is transferred from the veins to the arteries, that a pulse is made which may be heard in the chest."

The normal heart sounds are primarily due to the opening and closing of the heart valves. Various heart diseases which affect the valves can be diagnosed by typical changes in the heart sounds produced in this fashion. In addition, so-called *murmurs* can be detected in a variety of cardiac diseases such as mitral insufficiency or aortic insufficiency. These murmurs have different characteristics in the case of different disfunctions and thus are of specific diagnostic value.

6. **Supersonic waves** can be focused (see Chap. 30 on optical instruments regarding focusing) so that the sound energy is concentrated in a narrowly defined region. The waves may be made intense enough to destroy living tissues at the focus. This method has been used to destroy localized regions of the brain of an experimental animal without opening the skull. The destruction of the selected region is manifested by the inability of the animal to perform a certain function, such as, for instance, raising its left leg. Experiments as to the possibility of destroying cancerous tissues by this method have also been conducted.

QUESTIONS AND PROBLEMS

1. What is the difference between a musical tone and a noise?

2. In the battlefield supersonic whistles are used to call dogs. How would you proceed to detect such enemy dog calls without using elaborate equipment?

3. Do two sources of sound always produce a louder sound when sounded together rather than individually?

4. Severe explosions are known to have frequently caused automobile tires to burst. How do you explain it?

5. What does a violinist do in order to raise the pitch of a string when he tunes the instrument? Why does this procedure raise the pitch? How does he change the pitch when playing the instrument? Why does this process raise the pitch?

6. A violinist can produce flutelike sounds, the so-called *harmonics*, by placing his fingers *loosely* at certain points of the string. How do you explain the formation of this sound effect?

7. An organ pipe which is closed at one end emits a certain tone. What will happen to the pitch of this tone when the closed end is opened? What will happen to the pitch if the air temperature is raised? What will happen to the pitch when the air pressure is raised at constant temperature?

8. If you were to tune two violin strings to the same pitch, what criterion would you use to judge whether or not the pitch of both strings is exactly the same?

9. Could you suggest an explanation why the sound of a bell is much less pleasant than the sound of a piano string?

10. Imagine a child swinging in a swing with a large amplitude while continuously blowing a whistle of constant pitch. Describe what you would hear if you were to walk slowly around the swing and record your sound impressions at different places.

11. What is the difference between progressing waves and standing waves?

12. Explain the "rolling" of thunder.

13. If an organ pipe gives five beats per second against a tuning fork of 435 cycles/sec, what could you do to tell if the frequency of the pipe is 430 or 440 cycles/sec?

14. An organ pipe filled with air emits a tone of a certain pitch. What will happen to the pitch when the pipe is filled with CO_2 at the same pressure and temperature?

★**15.** Why is the tone color (timbre) of a closed pipe (closed at one end) different from that of an open pipe?

16. A tuning fork emits a sound of 100 cycles/sec. What pitch does an observer hear when he moves with the velocity of sound (*a*) toward the fork? (*b*) away from the fork? (*c*) past the fork?

17. Administration of male sex hormones to women lowers their voices. What changes in the speech organs could account for this change of voice?

18. List the ground tone and the overtones emitted by a pipe (*a*) open at both ends, (*b*) open at one end. Assume the pipe to be 60 cm long and the air temperature to be 20°C. (The ground tone and the overtones of this pipe are given by the *resonant frequencies* of a cylinder open (*a*) at both ends and (*b*) at one end.)

19. A shot is fired from a blimp which floats in the air 100 ft above a large lake. How large is the intensity of the echo of the shot heard on the blimp as compared with the intensity of the sound as heard in a boat directly under the blimp? Assume complete reflection of sound at the lake surface.

20. Compute the fundamental frequency and the second overtone of a string 40 cm long, of linear density of 0.3 gm/cm, stretched by a force of 10^6 dynes.

21. What should be the tension in a string 60 cm long having a linear density of 0.2 gm/cm if it is to emit a sound which gives rise to 10 beats per second against a 440-cycle/sec tuning fork? Is there a unique answer to this problem?

22. The intensity level of a sound of 10^{-10} watt/cm^2 intensity is increased by 4 decibels. Find the resulting intensity of sound.

23. A locomotive moving at the speed of 10 m/sec blows a whistle emitting 1,000 vibrations per second. What is the frequency heard by an observer located (*a*) 3 miles ahead of the train? (*b*) 2 miles behind the train? What frequency does a man on the train hear when the man in front of the train blows a whistle emitting 1,000 vibrations per second?

★**24.** An automobile is moving toward a wall. When a whistle of 300 cycles/sec is sounded, the echo of this sound (which is reflected by the wall) is heard by the observer in the car to produce three beats per second against the sound of the whistle. How fast is the car moving?

CHAPTER 28

THE NATURE OF LIGHT

a. How would you demonstrate that light is a form of energy?

b. What is meant by the term "linear propagation of light"?

c. What is the principle of the pinhole camera?

d. Define the terms luminous flux, intensity of a source, intensity of illumination.

e. What evidence can be cited in support of (1) the corpuscular theory, (2) the wave theory of light?

f. What objections can be cited against (1) the corpuscular, (2) the wave theory of light?

g. How can the processes of reflection and refraction be explained on the basis of (1) the wave theory, (2) the corpuscular theory of light?

h. What discrepancy is there in the predictions regarding the velocity of light between the corpuscular and the wave theory?

i. What experiment was considered a crucial one, demonstrating the classical corpuscular theory to be untenable?

j. What is "dispersion" of light?

k. What is the physical distinction between different colors of light?

l. Is it necessary to mix all colors of the spectrum in order to obtain white light?

m. What are complementary colors?

n. What methods for the determination of the velocity of light do you know?

Light as a Form of Energy

Our senses are sensitive detectors of different forms of energy. Just as a certain amount of energy is required to deflect the pointer of a meter, energy is supplied to a sense organ when it is stimulated. In the case of the auditory sense mechanical energy is supplied to the eardrum by the vibrating air particles. In the case of light a different kind of energy—we shall call it "radiant energy"—actuates tiny nerve endings in the retina of our eye.

To show that light is a form of energy we have merely to demonstrate that mechanical energy can be generated by the action of light. It is actually possible to impart rotational kinetic energy to a very light metallic paddle wheel W mounted on a nearly frictionless bearing in an evacuated tube (Fig. 28.1).

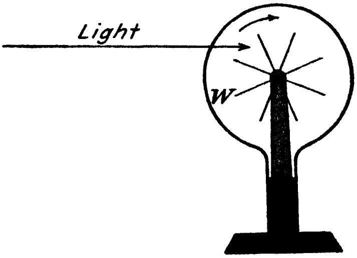


FIG. 28.1. Conversion of light energy into mechanical energy.

Light energy received from the sun is the ultimate source of the energy of the living organisms and of our present industrial power. Chlorophyll, the pigment of the green plants, enables them to synthesize carbohydrates from carbon dioxide and water under absorption of solar-light energy which is thus stored in form of the chemical energy of the synthesized organic compounds. Herbivorous animals derive their energy indirectly from the sun by eating these plants, and carnivorous animals feed on the solar energy even more indirectly by eating the herbivores.

Our industry depends on solar energy no less than our organism. The coal that feeds our industrial furnaces and the oil that feeds our engines are both products of disintegration of plants and animals which are derived from ancient cemeteries of prehistoric life. The remains of the ancient fauna and flora subjected to high pressure at an elevated temperature deep underground were profoundly changed in the course of ages, giving rise to coal and to oil. Thus we see that it is in the last analysis solar energy which we supply to our industrial plants, engines, and stomachs.

Linear Propagation of Light

Light energy is referred to as "radiant energy" because of its linear propagation. For instance, if we use a tiny luminous object *L* (*e.g.*, a small glowing sphere) as a "point" source of light and place before this source a wall *W* provided with a circular window (Fig. 28.2), we can determine the area of the screen *E* that will be illuminated by drawing straight lines *LC* and *LD*, which join *L* with the edges of the opening in *W*. Similarly, we can determine the region *S* (*shadow*) from which light will be blocked by the obstacle *B* by drawing the straight lines *LF* and *LG* tangent to the ball *B*.

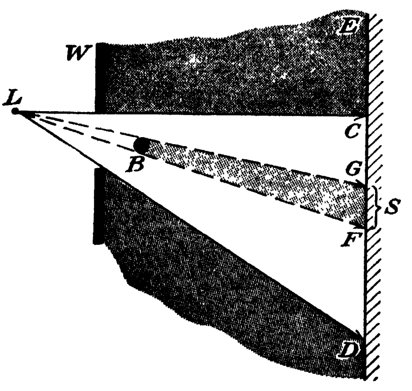


FIG. 28.2.

The linear propagation of light is utilized in the pinhole camera in the following fashion. Let us imagine a luminous object in the shape of an arrow, every point of which can be considered as a point source of light (Fig. 28.3). *C* is a box facing the arrow, *O* is a small opening in the front wall, and *G* is a ground-glass screen.

Let us consider first two representative luminous points *A* and *B*. The area of the screen *G*, which is illuminated by point *A*, is obtained by the same construction as shown in Fig. 28.2. The straight lines (rays) issuing from *A* toward the edges of the orifice *O* define a small, elliptical,† illuminated area *a* on the ground glass *G*. Similarly, light emanating from point *B* produces a tiny illuminated ellipse at *b*. Thus *for every luminous point of the arrow, we get a small, elliptical, bright patch on G.* As a result, an image of the arrow appears on the ground-glass screen. As the construction shows, *the image is inverted.* Owing to the overlapping of the elliptical images of individual points, the camera image is not sharp. As we make the hole *O* smaller, the elliptical patches become smaller and the image gets dimmer but sharper.

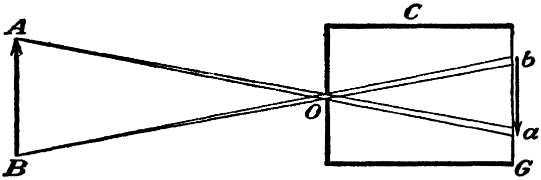


FIG. 28.3. Image formation in a pinhole camera.

The pinhole camera can be used with radiations such as X rays, which cannot be focused by a lens as light is in a photographic camera. The pinhole camera has also been used to obtain natural-color photographs of the interior of the stomach for ulcer diagnosis. The camera has to be swallowed by the patient and, hence, must be very small. A pinhole camera is used for the purpose, since no lenses of the required small dimensions capable of producing sharp pictures in natural colors are available.

The Inverse Square Law

The energy from a point source of light emanating uniformly in all directions is "diluted" in a similar fashion to sound energy as it recedes from the source. The same considerations which led to equations (27.1) and (27.2) result in the derivation of the **inverse square law for the intensity of illumination**: If Φ is the **luminous flux** of the source, *i.e.*, the total amount of light energy radiated from it in all directions per second, $\Phi/4\pi R^2$ is the amount of energy falling upon each square centimeter of a sphere of radius R which surrounds the source. We shall define the energy impinging upon unit area as **intensity of illumination** E. The illumination at any distance R is then given by

$$\boxed{E = \frac{\Phi}{4\pi R^2} = \frac{\Phi}{A}} \qquad (28.1)$$

---

† The ellipse is produced by oblique intersection between the light cone and the plane of the glass screen *G*.

The equation $E = \Phi/A$ is our *definition equation for the intensity of illumination* (or illumination for short). A is the illuminated area, and Φ is the **luminous flux** impinging upon it. We see that the **illumination is inversely proportional to the square of the distance between the illuminated area and the source.** For the unit distance of $R = 1$ cm, the intensity of illumination becomes

$$E_0 = \frac{\Phi}{4\pi} = I$$

The expression $\Phi/4\pi = I$ is called the **luminous intensity of the source.**

Thus we can write equation (28.1) in terms of the luminous intensity as follows:

$$\boxed{E = \frac{I}{R^2}} \tag{28.2}$$

When applied to actual sources of light this equation must be considered an approximate law, since the assumptions made in its derivation are not exactly fulfilled by practical sources. (1) No practical source is a true point source. *Any source, however, can be considered approximately a point source if its size can be neglected as compared with its distance from the illuminated area.* (2) A source of light may not radiate uniformly in all directions. In this case E is not a function of R alone.

Early Ideas on the Nature of Light

To a contemporary student the idea that light is a form of energy and that this energy travels from a luminous point to our eye where it stimulates visual nerve endings appears quite natural. But our modern ideas on the process of vision are not at all self-evident as can be seen from the fact that thinkers of the stature of Pythagoras, Hipparch, Euclid, and Ptolemy conceived of the process of seeing as being due to a radiation emanating from the pupil of our eye, the rays serving us as a sort of feeler similar to the stick of a blind man. These feelers must be very long indeed when we look at a distant star! To Aristotle's embarrassing question: "If the eye is of a fiery nature, why don't we see in the dark as well?" Plato found the following answer: He assumed that the eye emits an "inner light" as long as it is kept open but that these rays are not effective in making us see the objects unless there is present at the same time an "outer light" such as the sun, moon, or a candle. Aristotle expressed views on the nature of light which came close to the ideas of the nineteenth century. He assumed vision to be due to interaction between the eye and the perceived body, this interaction being effected by the intermediary "ether," the fifth element of the ancient Greeks, a transparent all-permeating substance through which the sun and the stars were supposed to be moving.

The Corpuscular and the Wave Theory of Light

The speculations of the ancients were restricted by the fact that they could not think of light except in connection with the process of seeing. The realization that light is something which exists quite independently of our eye, the conception of light according to which it could be present in a world uninhabited by living creatures, meant a decisive advance beyond the speculations of the ancient philosophers. We have come eventually to the realization that emission of light is a transfer of energy. But through what mechanism is this energy transfer accomplished? Until the close of the nineteenth century, most physicists felt that a physical phenomenon was not understood unless a mechanical picture describing the process was provided. There are essentially two mechanical ways of transmitting energy. There is *convection*, involving transport of matter (transfer of kinetic energy of particles) from the source to the observer, and there is *wave motion*, which can transmit energy without transport of matter (*e.g.*, sound) from the source to the observer. The latter process depends on the presence of a medium between the source and the observer, which serves as the carrier of the wave motion.

Thus two mechanical theories of light have reached particular prominence, the **corpuscular theory** and the **wave theory.** The linear propagation of light (as revealed by the study of the formation of shadows) suggests quite naturally that light may consist of tiny corpuscles moving with uniform speed in straight lines according to the law of inertia. Newton favored this view more and more in the later years of his life. This view would explain without additional implausible hypotheses why light can travel through empty space (produced with the aid of a vacuum pump in the laboratory or encountered in the vast region separating the earth from the stars and planets). It has been also argued, however, that light may be just as well due to a motion of corpuscles which resembles "motion of the particles in a sound wave rather than the motion of air particles in a wind." In other words, there may be transfer of mechanical energy through a wave which does not involve the transfer of corpuscles from the source to the illuminated body. **This view requires the hypothesis that all space (vacuum as well as the interior of bodies) is filled with a subtle medium, which we shall call the *ether*, consisting of tiny particles which may transfer energy to their neighbors by elastic interaction. Waves which are engendered in this medium are assumed to constitute light.**

One of the arguments advanced by Huygens in support of the wave theory was the fact that two light beams cross each other without disturbing each other's path. If light consisted of particles, he argued, some of them would collide and light would be scattered in all directions from the region of intersection of the two beams. Nothing of that sort is observed, however.

The corpuscular theory, in which we do not have to assume the existence of a "luminiferous ether," is, at first glance, undoubtedly simpler than the wave theory. But Newton soon realized that the theory had to be made more complicated in order to explain more than just the linear propagation of light. He knew that part of the light striking a glass plate is reflected while part of the light goes through. Why should some of the light corpuscles be reflected and others transmitted? What is the difference between the reflected and the transmitted corpuscles? To explain this, Newton was compelled to ascribe to his corpuscles periodic properties making them resemble vibrating drops rather than rigid particles. The state of a light corpuscle was supposed to fluctuate, assuming periodically "fits of easy transmission" and "fits of easy reflection." We find in his "Opticks," Book II, Prop. 12, the following passage:

> . . . every ray of light, in its passage through any refracting surface is put into a certain transient constitution or state, which, in the progress of the ray, returns at equal intervals, and disposes the ray, at every return, to be easily transmitted through the next refracting surface, and between the returns, to be easily reflected by it.

Newton retained the ether as an intermediary in the process of interaction between light particles and matter. Ether was supposed to be in different states of condensation in different kinds of matter. Corpuscles encountering a region of varying ether density would be deviated from their path, exhibiting reflection or refraction.

We shall examine in the following sections how the various phenomena exhibited by light can be explained by both theories of light in an attempt to decide which one of these two theories is preferable to the other. Whereas the corpuscular theory offers a very simple explanation for the linear propagation of light, the wave theory proved initially incapable of doing so, and this failure was considered by Newton as a good reason for its rejection. If *L* in Fig. 28.2 were a source of sound instead of a source of light, observers stationed above *C* and below *D* would hear the sound. Sound travels, so to speak, "around the corner." The opponents of the wave theory argued that according to the wave theory of light regions above *C* and below *D* should be illuminated. We shall see in Chap. 29 that this prediction has actually been confirmed, but the bending of light around the corner is so slight that it was initially disregarded. The smallness of this effect in the case of light as compared with observations made with sound was satisfactorily explained by the wave theory as will be shown in Chap. 29.

The Corpuscular Theory and the Wave Theory of the Reflection of Light

Reflection of light can be explained equally well on the basis of either theory. If we allow a beam of sunlight to enter through a small hole in a window shade into a dark, dusty room, the dust particles are illuminated along the path of light and we see a narrow beam of light. In Fig. 28.4 the beam of light AO is seen to impinge upon a mirror (a polished surface) MM. After reaching the mirror, light proceeds in a different direction, namely, OB; it has been reflected. Measurements show that the **angle of incidence i is equal to the angle of reflection r.** These angles are measured between the light ray and the normal N (*i.e.*, the perpendicular to the surface) at the point of incidence. The corpuscular theory explains this law as follows: Light corpuscles behave like elastic balls. The momentum of an elastic ball moving obliquely toward a reflecting elastic plane surface can be resolved into two components: $(mv)_x$ parallel to the plane and $(mv)_y$ perpendicular to it (Fig. 28.5). The parallel component is not affected by the process of reflection: $(mv)'_x = (mv)_x$, but the perpendicular momentum vector is reversed in direction: $(mv)'_y = -(mv)_y$. It is evident from the diagram that angles α and α' formed by the initial and final momentum vectors with the normal N are equal.

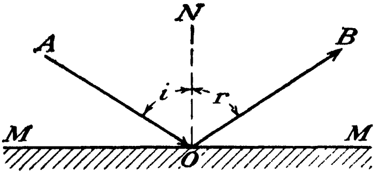


FIG. 28.4. Reflection of a light ray.

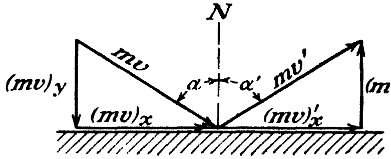


FIG. 28.5.

The process of reflection as interpreted by the wave theory is illustrated in Fig. 26.14. The vector of wave velocity (the "propagation vector") is perpendicular to the wave front. It can be easily shown that the angle between the wave front AB and the reflecting surface R_1R_2 is equal to the angle between v and the normal. (The legs of these two angles are mutually perpendicular.) Therefore, the equality of the angles i and r in Fig. 26.14 implied the equality of the angles of incidence and reflection as measured between the velocity vectors and the normal.

The latter angles are identical with i and r in Fig. 28.4. Thus, both theories predict the same law of reflection.

Wave Theory of the Refraction of Light

The explanation of the refraction of waves was discussed in Chap. 26. Figure 26.15 illustrates how the direction of a wave front AB which forms the angle i with the plane R_1R_2 separating two media (in which the wave velocity is different) is altered when it passes into the second medium. It is evident from the diagram that the angle through which the "propagation vector" has turned (compare vectors v_1 and v_2) is the same as the angle through which the wave front has rotated.

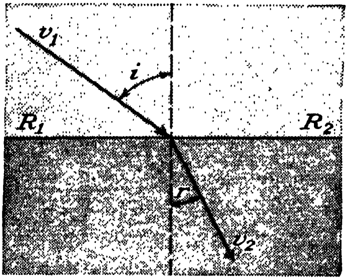


FIG. 28.6. Refraction of a light ray.

As in the case of Fig. 26.14 illustrating reflection, it can be shown here that the angle between a "propagation vector" v and the normal N (Fig. 26.15) is the same as between the corresponding wave front and R_1R_2. In other words, the angles i and r in Fig. 28.6 are the same as i and r in Fig. 26.15 and the relation (**the law of refraction**)

$$\boxed{\frac{\sin i}{\sin r} = n = \frac{v_1}{v_2}} \tag{28.3}$$

holds equally well for the diagram of Fig. 28.6.

According to the interpretation of the wave theory, the velocity v_2 in the second medium must be smaller than v_1 if the ray is refracted toward the normal, i.e., if $r < i$. Measurements show that the indices of refraction of all common optical media with respect to a vacuum ($n = c/v$)† are larger than 1 for visible light. This means that *the velocity of light waves in these material media must be slower than in a vacuum according to the wave theory.* The index of refraction of gases is very close to 1 suggesting that light is inappreciably slowed down in gases.

Dispersion and Colors

Imagine the experiment indicated in Fig. 28.6 to be carried out with light of different colors. We admit a thin ray of sunlight represented by the arrow v_1 through a small opening in a window shade. The beam of light is refracted at the interface R_1R_2 between the two media; *i.e.*, its direction changes abruptly so that it proceeds in the direction of the arrow v_2. In order to obtain a beam of light of a desired color we place a "filter" of colored glass in front of the opening in the shade. We notice that, while the angle i is kept constant, the angle r varies as the color of the incident light is changed. In other words, *light of different colors has different indices of refraction. Blue light is refracted more strongly than red light; it has, according to our definition, a larger index of refraction,* the angle r being smaller for blue light than for red light. If we perform the experiment shown in Fig. 28.6 with vacuum as medium 1 and, say, glass as medium 2, using successively red, yellow, green, blue, and violet rays of light all moving along the path indicated by v_1, we get five different refracted rays shown in Fig. 28.7. If we manage to pass those rays simultaneously down the path v_1, the five refracted rays will be seen at the same time. The angular separation between the rays is greatly exaggerated in Fig. 28.7. In order to separate the end points of these rays more widely, they should be made as long as possible. To avoid the use of an excessively large body of glass, one can use a prismatic glass block as shown in cross section in Fig. 28.8. Light is refracted at both surfaces S_1 and S_2, and the two rays R and V, after moving a long distance through air, produce two widely separated patches of light on a distant screen.

---

† c designates the velocity of light in a vacuum.

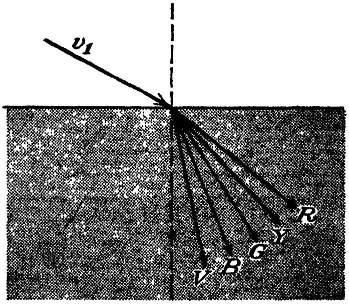


FIG. 28.7. Illustration of dispersion. (The deviation of the rays is greatly exaggerated.)

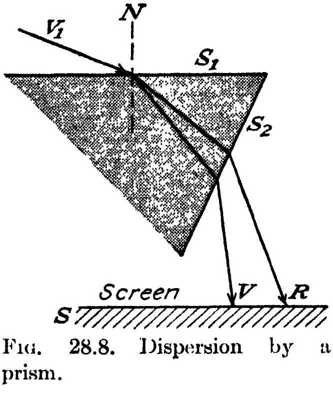


FIG. 28.8. Dispersion by a prism.

Newton sent a beam of sunlight through a prism and observed that the luminous patch produced on a distant screen by the light transmitted through the prism was multicolored. A continuous colored band (spectrum) extended from the red edge (R in Fig. 28.8) to the violet edge (Fig. 28.9). Newton explained his observation by assuming that white light is a mixture of light of different colors and that this mixture is decomposed by a prism because the different colored components of white light are not refracted equally strongly. He expresses his ideas on the nature of colors as follows in the *Philosophical Transactions* 80 (1671–1672):

> Colours are not classifications of light derived from Refractions, or Reflections of natural Bodies (as 'tis generally believed), but original and connate properties, which in divers rays are divers. Some rays are disposed to exhibit a red colour and no other: some a yellow and no other, some a green and no other, and so of the rest. Nor are there only rays proper and particular to the more eminent colours, but even to all intermediate gradations.
>
> To the same degree of Refrangibility ever belongs the same colour, and to the same colour ever belongs the same degree of Refrangibility.

Newton likened the different colors of light to sounds of different pitch, and it was perhaps this acoustical analogy which induced him to distinguish seven different colors of light in the sun's spectrum, there being seven tones in our conventional musical scale. The spectrum is arbitrarily subdivided into the following colors: red, orange, yellow, green, light blue, dark blue (indigo), violet.

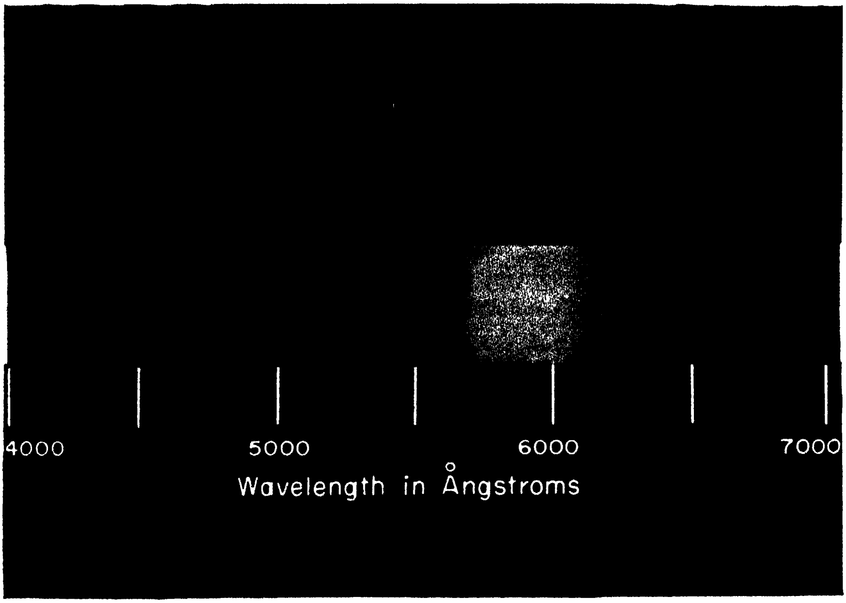


FIG. 28.9. A continuous prism spectrum. (*From Weber, White, and Manning, "College Technical Physics," McGraw-Hill.*)

In addition to demonstrating that white sunlight is a mixture of different colors, Newton showed that, conversely, by mixing light of these "spectral" colors we can produce white light. It can be shown, however, that not necessarily all the spectral colors are required to produce white light. A red and a green beam of light mixed in a suitable proportion of intensities illuminating the same area will make it appear white. The same can be said of the color combination of blue and yellow and of a variety of other pairs of colors. Colors which combine to give the sensation of white are called **complementary colors.**

Complementary colors are also encountered in the following experiment. We can produce a spectrum of white light as indicated in Fig. 28.8. Imagine the large screen S to have been replaced by a very narrow strip, which obstructs the passage of red light only. It is possible to unite the remaining rays again and make them illuminate an area. (This can be done by means of an oppositely oriented prism.) It is observed then that the area illuminated by the residual colors appears blue-green. In other words, it shows the color complementary to red. Similarly, elimination of the blue light from the spectrum leads to a yellow ray when the remaining colors are combined. In general *the color of the residual combined light is complementary to the removed light.*

What is the physical difference between green, red, and blue light? *According to the wave theory,* a source of light is analogous to a source of sound. It generates vibration of ether particles, which give rise to wave motion in the ether. If we assume the different colors of light to correspond to vibrations of different frequency, we should expect the wave lengths of light corresponding to different colors to be different. But we have no way of deciding at the moment whether the higher frequency of light should correspond to the red or to the blue light.

According to the corpuscular theory of light, we shall imagine that different kinds of corpuscles constitute light of different colors. Newton's explanation of colors combined features of the corpuscular theory with those of the wave theory (periodicity). As we have seen above, the periodicity feature was introduced by his hypothesis of periodic variation of the state of a light particle between "fits" of easy reflection and "fits" of easy transmission. According to this hypothesis the color depends on the duration of the interval between two successive repetitions of the same "fit."

The Corpuscular Theory of the Refraction of Light

We have explained the law of refraction of light on the basis of the wave theory, from which the law of refraction $(\sin i)/(\sin r) = n$ can be derived. Is it also possible to deduce the correct law of refraction on the basis of the corpuscular theory? What assumptions must be made in order to get a result which agrees with experiment?

Let us assume that the light corpuscles are attracted toward the surface of a transparent medium as they enter the medium coming from empty space. The force of attraction is perpendicular to the surface. A light corpuscle moving obliquely toward the surface with velocity v_1 (Fig. 28.10) will be accelerated toward the surface and will enter the material medium with a higher velocity v_2. We can decompose the velocity vectors v_1 and v_2 into a component parallel to the surface (v_x) and a normal component (v_y). The v_x component remains unchanged in the absence of forces parallel to the surface. The change occurs only in the v_y component. This consideration is valid not only when the first medium is a vacuum but in general for any two media which do not attract the light corpuscles equally strongly.

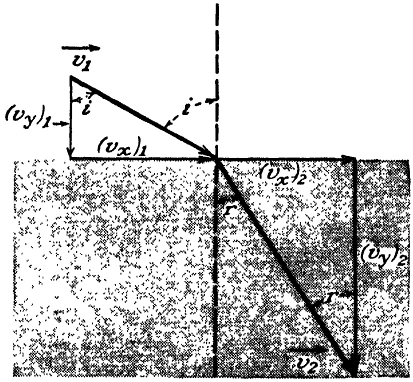


FIG. 28.10.

We can read off the following relations from Fig. 28.10:

$$\sin i = \frac{(v_x)_1}{v_1}$$

and

$$\sin r = \frac{(v_x)_2}{v_2}$$

Since the x component of v does not change, we can write

$$(v_x)_1 = (v_x)_2 = v_x$$

hence,

$$\sin i = \frac{v_x}{v_1} \quad \text{and} \quad \sin r = \frac{v_x}{v_2}$$

Consequently,

$$\frac{\sin i}{\sin r} = \frac{\cancel{v_x}\, v_2}{v_1 \cancel{v_x}} = \frac{v_2}{v_1} = n^* \tag{28.4}$$

This equation derived from the corpuscular theory is of the same type as equation (28.3) derived from the wave theory. The ratio of the sines of the angles of incidence and of refraction is equal to a constant (refractive index n^*), which is equal to the ratio of the velocities of propagation of light in the two media. *There is, however, an important difference between equations* (28.3) *and* (28.4). Let us write these equations for simplicity of comparison, assuming vacuum to be the first medium; then $v_1 = c$, and we get

$$\frac{\sin i}{\sin r} = \frac{c}{v_2} \tag{28.3'}$$

and

$$\frac{\sin i}{\sin r} = \frac{v_2}{c} \tag{28.4'}$$

We know from experience that in all cases $i > r$ if i is the angle in the vacuum. Hence, the first equation states

$$\frac{c}{v_2} > 1 \quad \text{or} \quad c > v_2 \tag{28.3''}$$

and the second one, $$\frac{v_2}{c} > 1 \quad \text{or} \quad v_2 > c \qquad (28.4'')$$

Thus we see that, although both the wave theory and the corpuscular theory yield a law of refraction of the same form, they lead to *contradictory predictions about the velocity of light in material media. According to the wave theory* (28.3″) *the velocity of light in matter is smaller than in a vacuum, whereas according to the corpuscular theory* (28.4″) *it is larger than in a vacuum.*

Velocity of Light

This result seems to suggest that one of the two theories must be wrong. The method of determining which one is the correct theory appears to be obvious: Measure the velocity of propagation of light in various media and compare your data with the velocity of light in a vacuum. This program is, however, far from easy to carry out. It was suspected long ago that light may require a finite time to travel from its source to its destination. Galileo suggested the following method to measure the velocity of light, which was actually attempted by the Florentine Academy.

Galileo's Method. An observer A is equipped with a timing device (sand clock) and a well-shielded lantern, which is furnished with a shutter that can be opened to allow light to escape. A few miles from observer A is stationed a man B with a similar lantern. A opens the shutter and simultaneously starts his timing device. As soon as the light signal reaches B, he opens his shutter, which sends a light signal toward A. As soon as A sees the light flash released by B, he stops the timer and reads the time elapsed. The knowledge of the time interval and the distance between A and B should have yielded the velocity of light. These experiments showed only that the velocity of light is either infinite or far too great to be determined by such crude means.

Roemer's Method. As in many other instances the solution of this physical problem emerged from astronomical studies. In Galileo's days one did not have the precise chronometers giving the Greenwich time on board a ship for the determination of the geographic longitude. Galileo made the ingenious suggestion that one of the Jupiter moons, which he discovered with his telescope, could be used as a celestial clock for purposes of navigation. The satellite circles around its orbit like the pointer of a clock, and the completion of one revolution is easily recognized by watching the satellite enter Jupiter's shadow and then disappear. To calibrate this celestial clock for practical purposes would involve the computation of the sidereal times of onset of the eclipse of the Jupiter moon. Cassini, the director of the Paris observatory, became interested in this problem. His assistant and collaborator, Olaf Roemer, was particularly successful in collecting and interpreting the necessary data. We shall present Roemer's observations with some crude simplifications in order to avoid a discussion of technical astronomical details. We shall disregard the motion of Jupiter in the time interval of a terrestrial half year and shall furthermore make the assumption that observations of the Jupiter moon can be carried out during the daytime as well as at night. Let us pretend that Roemer carried out his observations under the following conditions. While the earth was in position A (Fig. 28.11), he observed the period of the satellite by measuring the time between two consecutive eclipses of this Jupiter moon. He was then able to predict the exact moment when the hundredth eclipse should take place about half a year later, *i.e.*, when the earth was in position C.

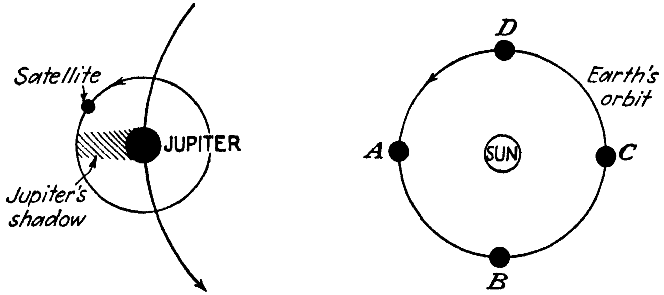


Fig. 28.11. Illustration relating to Roemer's determination of the velocity of light.

Much to his surprise the eclipse took place about 16 min later than he predicted. He then immediately checked† the period which he determined while the earth was at A and found his original value confirmed. He predicted again the time of occurrence of the hundredth eclipse half a year later, when the earth has returned to position A. This time the eclipse took place 16 min earlier than he predicted. These observations could be explained as variations in the period of revolution of the *Jupiter moon,* assuming the period to be longer than the average period when the earth moves away from Jupiter and shorter when it moves toward it. Such a relation between the period of the satellite and its motion relative to the earth looked very improbable to Roemer. He preferred to explain the phenomenon by assuming a finite velocity of propagation of light. Roemer's argument could be presented as follows: If we had made our measurements at A (Fig. 28.11) and had left at the point A of the earth's orbit an assistant who would remain there without participating in the earth's revolution about the sun, this assistant would have observed the eclipse of the Jupiter moon precisely at the predicted moment. At that time the earth would be at the point C. If we assume that light travels with a finite speed, we should expect the signal of the beginning eclipse to reach us at C at a later time than it reached our assistant at A. The discrepancy between the computed and observed time of the eclipse can be considered as being equal to the delay in the observation of the eclipse due to the fact that light has to traverse the diameter of the earth's orbit in traveling from the assistant at A to our position at C. If we assume this explanation to be correct, we can compute the velocity of light in the interplanetary vacuum as follows:

---

† This account of Roemer's procedure is fictitious and has been used for the sake of clarity and simplicity.

$$c = \frac{\text{diameter of the earth's orbit}}{\text{discrepancy between predicted and observed time}}$$

From the data available to him, Roemer found a value for the velocity of light which is very close to modern determinations. c was found to have the very large value of **3×10^{10} cm/sec,** which explains Galileo's lack of success in measuring it by his crude method.

Foucault's Method. In order to decide between the corpuscular theory and the wave theory of light physicists had to devise a method which would permit the determination of the velocity of light in various material media as well as in a vacuum. Foucault developed a laboratory setup, which has been later modified and greatly improved by Michelson.

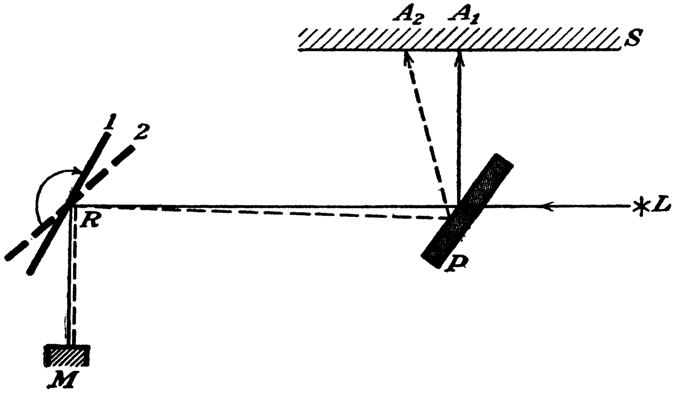


FIG. 28.12. Foucault's method of determination of the velocity of light.

Foucault's apparatus is shown schematically in Fig. 28.12. A beam of light emanates from the light source L, passes through the glass plate P (refraction in the glass plate has been disregarded in the drawing), and impinges upon the mirror R which is, to begin with, in position 1. It is then reflected toward the very narrow mirror M, is reflected from M back to R, is again reflected from the ("half silvered") left surface of the plate P, and finally gives rise to the bright spot A_1 on the screen S. This light path is indicated by solid lines in the diagram.

When the mirror R is rotated slightly and slowly out of position 1, the beam reflected from R does not hit M any more and the bright spot A_1 disappears. (This is *not* the situation pictured by dotted lines in Fig. 28.12.) As the mirror rotates slowly, the spot A_1 flickers faster and faster as the rotational speed of the mirror increases. When the frequency of revolution surpasses 16 per second, the flickering sensation disappears and we see again a stationary spot at A_1. But if the frequency of revolution is increased very much more, the spot A begins to shift with increasing angular velocity of the rotating mirror R to the position A_2. This shift can be explained as follows: After reaching the small mirror M with R in position 1, the light signal is reflected and follows the dotted path. At very high rotational speeds, the mirror R has enough time to rotate through an appreciable angle from position 1 to position 2 while the signal travels from R to M and back to R. When the reflected light arrives at R on its way from M, it finds R in position 2 and does not retrace the solid path but rather moves according to the laws of reflection, following the dotted path which is determined by the new position of the mirror. As a result, light is reflected toward A_2. From the shift A_1A_2 we can determine the magnitude of the angle separating the positions 1 and 2 of the rotating mirror. Knowing the frequency of revolution, we know the time required for one revolution and, hence, the fractional time required for the rotation from position 1 to position 2. Let us call this time t. During this time the light signal has moved from R to M and back, *i.e.*, the distance $2MR$; hence, the velocity of light is

$$c = \frac{2MR}{t}$$

This method yielded a value for c in a vacuum in good agreement with Roemer's result.

Foucault used a frequency of revolution of 800 revolutions per second. The distance RM in his apparatus was 10 m. This setup was on a small enough scale to enable him to interpose water and other materials between R and M and thus to determine the speed of light in material media.

Foucault found that the velocity of light in all material media which he tested was smaller than in a vacuum. This was considered a crucial experiment deciding in favor of the wave theory of light and rejecting the corpuscular theory as untenable. As we shall see later, this conclusion turned out to be correct in a limited sense only. Foucault's experiment merely ruled out "Newtonian corpuscles," which move according to the laws of Newtonian mechanics and which are speeded up as they enter a denser medium. There still remains, however, a possibility that light corpuscles obey different laws of physics. This question will be taken up in connection with the photon theory of light in Chap. 35.

Foucault also found that red light travels faster than blue light in, say, glass or water. This finding agrees with the prediction of the wave theory of light according to which [see equation (28.3)] the waves that are refracted more strongly (blue) should be the slower ones.

While light of different colors has been shown to travel with different speed in material media, it is known to travel with the same speed in a vacuum irrespective of color. This conclusion follows from simple evidence. Occasionally we observe the birth of a star. A "supernova" suddenly flares up among the stars at enormous distances from our earth. Light from these stars may take millions of years to reach us. If, say, red light were to travel faster than light of other colors, we should expect the new star to appear red at first and only after a considerable delay to acquire its final white appearance. Other evidence is provided by double stars ("eclipsing binaries"). These are systems of two stars, one of which is faint. They revolve about a common mass center and are sometimes so disposed with respect to us that the faint star periodically eclipses the bright one. If light of different colors were to travel with different speeds, we should not see a white star "flash on and off" but rather we should observe the star to change its color before appearing white and before disappearing. Such phenomena have never been observed. All colors of the spectrum arrive at the same time, which shows that *there is no dispersion in a vacuum.*

APPENDIX

Photometric Units

Luminous Intensity. One unit of luminous intensity is *the international candle.* It is defined arbitrarily by specifying the design and operating conditions of an electric standard lamp whose luminous intensity is by definition one international candle.

Luminous Flux. The unit of luminous flux is one *lumen.* It is the luminous flux passing through a unit area on a unit sphere (a sphere of unit radius) described about a point source of 1 international candle.

Intensity of Illumination. The unit of illumination is 1 *foot-candle* (*or* 1 *meter-candle*). It is the illumination of the inner surface of a sphere of 1 ft (or 1 m) radius described about a point source of 1 international candle. *At this illumination the flux per unit area is* 1 *lumen/ft*^{2} (*or* 1 *lumen/m*^{2}).

QUESTIONS AND PROBLEMS

1. To what kinds of processes do you expect the inverse square law to be applicable besides illumination by visible light?

2. Can you demonstrate dispersion of light without using a prism?

3. Can you produce, experimentally, conditions where light will travel along a curved path?

★4. What effect do you expect the revolution of the earth about the sun to have upon the observed position of stars. Assume the stars to be so distant that the change of the earth's position in its orbit would not produce a noticeable change in the observed relative position of the stars.

5. What is the difference between scattering and reflection of light? Give examples.

6. If you look for a long time at a red spot and then quickly fix your eyes on a white surface, you have the illusion of seeing a green spot on the blank white surface. How do you explain this?

7. When a beam of light coming from under the water surface hits the surface, part of it is generally reflected and part of it is transmitted. At a certain angle (critical angle of *total reflection*) the transmitted beam disappears and only the reflected beam persists. Explain this phenomenon, and compute the value of the critical angle from the known refractive index of water.

8. *a.* Construct the shadow of a sphere of 1 cm radius which is placed midway between a point source of light and a wall which is 20 cm away from the light source. How does the size of the shadow compare with the size of the object?

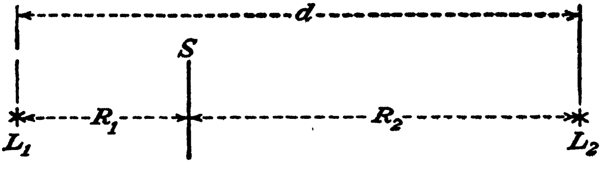
b. Assume a second point source placed 1 cm to the right of the first source. Draw the shadow under these conditions. In what way has the original shadow been modified by the presence of the second light source.

9. If the shadow of a cross consisting of two straws is projected on a wall by means of a candle, the shadow of the vertical rod of the cross is much sharper than the shadow of the horizontal rod. Why?

10. The image of a house is produced on the ground glass of a pinhole camera. The pinhole is a circle 1 mm in diameter. How will the image of the house be changed when half of the pinhole is covered up?

11. A vertical beam of light impinges upon a horizontal mirror. Find the direction of the reflected beam after the mirror has been rotated 10° about a horizontal axis. What angle is swept by the reflected light beam when the angle of rotation of the mirror is increased from 10 to 12°?

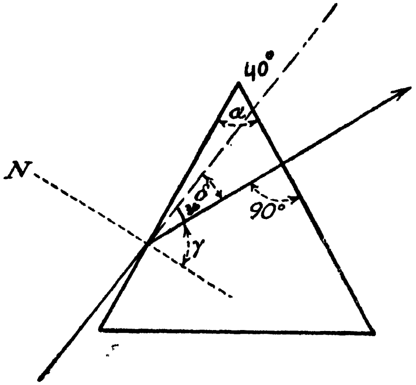
12. The figure represents the scheme of a *photometer*, a device for the comparison of the intensity of two light sources. The intensity of the source L_1 is known to be $I_1 = 20$ candles, and the intensity of the source L_2 is to be determined. S is a white screen illuminated from both sides by L_1 and L_2. This position between L_1 and L_2 is adjusted so that both sides of S are illuminated equally. A mirror arrangement, not shown in this figure, enables an observer to see both sides of S simultaneously next to each other and thus to compare their illumination. How large is the "candle power" of L_2 if both sides of S are equally bright for $R_1 = 20$ cm? Assume as distance between the two sources $d = 50$ cm.



13. The illumination of a book page is 20 foot-candles. How far should one remove the page from the illuminating point source in order to reduce the illumination to 5 foot-candles? How large is the light flux striking the page if its dimensions are 8 by 12 in.?

14. A beam of light impinges upon the water surface forming the angle $\alpha = 30°$ with the normal. Find the deviation from the normal of the light beam under water. Take as index of refraction of water $n = 1.33$.

★15. A light ray enters a prism whose apex angle is $\alpha = 40°$. The beam is refracted so as to enter the second surface at right angles. If the deviation of the beam is $\delta = 30°$, find the refractive index.



16. How large must be the angle between a light beam traveling under water toward the surface and the normal if the emerging beam is to be parallel to the water surface? What will happen if the angle under water is increased beyond the above value?

17. How large is the speed of yellow light in crown glass if its index of refraction is $n_y = 1.517$? How large is the speed of blue light in crown glass? ($n_b = 1.523$.)

18. A ray of light aiming toward the surface of a flint glass slab forms an angle of 30° with the normal inside the glass. The glass slab is under water. What is the direction of the ray which emerges into the water? (Take for flint glass $n' = 1.58$ and for water $n'' = 1.33$.)

19. What would be the direction of the emerging ray in Prob. 18 if we were to replace the water by an oil of an index of refraction equal to that of the glass?

CHAPTER 29

LIGHT AS A WAVE MOTION

a. What do you consider the most convincing evidence in support of the wave theory of light?

b. Is it possible to produce standing light waves?

c. How can the wave length of light be measured?

d. What physical characteristics of light determine its color?

e. What is a diffraction grating?

f. What do we mean by the term "order of interference"?

g. Can light "bend around a corner"?

h. Can you give an example of diffraction of light?

i. Why do soap bubbles appear colored?

j. How are Newton's rings produced?

k. When will a glass plate appear dark in reflected monochromatic light?

l. Why is the center in the pattern of Newton's rings black when viewed by reflected light? Is it black only for a particular color of illuminating light?

m. Are there only seven spectral colors?

n. What is the advantage of a diffraction grating over a double slit?

o. Is it possible to determine the wave length of light using a single slit?

Interference of Light

The wave theory of light was considered as rather securely established long before the performance of Foucault's "crucial experiment." This was in a large measure due to the pioneer work of the English physician Thomas Young. About 1800 he wrote:†

> Suppose a number of equal waves of water to move upon the surface of a stagnant lake, with a certain constant velocity, and to enter a narrow channel leading out of the lake; suppose then another similar cause to have excited another equal series of waves, which arrive at the same channel, with the same velocity, and at the same time with the first. Neither series of waves will destroy the other, but their effects will be combined; if they enter the channel in such a manner that the elevations of one series coincide with those of the other, they must together produce a series of greater elevations; but if the elevations of one series are so situated as to correspond to the depressions of the other, they must exactly fill up those depressions, and the surface of the water must remain smooth. Now I maintain that similar effects take place whenever two portions of light are thus mixed; and this I call the general law of the **interference of light.** Thus, whenever two portions of the same light arrive to the eye by different routes, either exactly or very nearly in the same direction, the light becomes most intense when the difference of the routes is any multiple of a certain length, and least intense in the intermediate state of the interfering portions; and this length is different for light of different colors.

We recognize that Young makes reference to the phenomenon of interference of waves, which was discussed in Chap. 27. We saw how a periodic pattern is established in space, nodal regions (zones of quiescence) alternating with antinodes. Such a pattern at once reveals the wave nature of a process and allows the determination of the wave length through the measurement of the distance between two neighboring nodes. For instance, when sound waves emanating from a continuously sounding source are reflected from a wall, the interference between the direct and the reflected sound wave sets up a pattern of zones of silence and zones of maximum loudness, which demonstrates the wave nature of sound and reveals its wave length. A similar experiment can be performed with *two* sources of sound or two sources of water waves of equal frequency. Figure 29.1 shows a photograph of the standing-wave pattern, observed on the water surface, which is agitated by two "sources" S_1 and S_2, represented by two spheres vibrating up and down in phase. Figure 29.2 shows a schematic diagram of the same phenomenon. The solid circles represent crests and the dashed ones troughs of the waves. Points of intersection where the two waves always meet in phase (crest and crest or trough and trough) are marked by small circles. These points are characterized by the fact that their distances from S_1 and S_2, respectively, differ by an integral multiple of a wave length.† At these points the waves from S_1 and S_2 arrive in phase; hence, these are points of permanent maximum agitation. The small squares, on the other hand, mark points where the two waves always meet in phase opposition. These are points of permanent minimum agitation, so-called **nodal points.** *Nodal points are characterized by the fact that their distances from S_1 and S_2 differ by an odd number of half wave lengths* as can be seen from the figure. Solid lines (hyperbolas) joining the circles mark *antinodal regions*, whereas

---

† Young's Works I, p. 202.

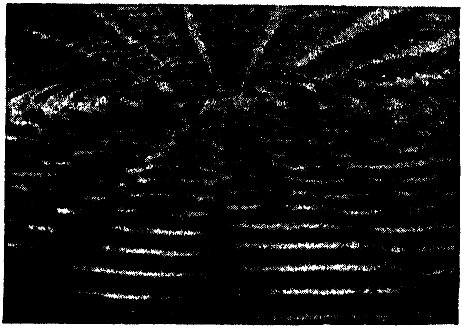


FIG. 29.1. Photograph of interference pattern of surface waves. (*From Freeman, "Modern Introductory Physics," McGraw-Hill.*)

---

† For example, the distance of P_1 from S_1 in Fig. 29.2 is $4\frac{1}{2}\lambda$, where λ is the wave length as measured by the distance between two consecutive solid circles. The distance of P_1 from S_2 is $3\frac{1}{2}\lambda$. Thus the path difference $(P_1S_1) - (P_1S_2) = \lambda$.

dashed lines joining the squares mark the *nodal regions* of relative quiescence. In the case of three-dimensional propagation in space we get nodal and antinodal *surfaces* rather than *lines*.

A similar experiment can be performed with sound. If S_1 and S_2 are the tips of the "tines" of a vibrating tuning fork, we obtain a pattern of interference of sound waves similar to the one shown for water waves in Fig. 29.2. There are zones of maximum loudness and of minimum loudness. This can be demonstrated simply by holding a vibrating tuning fork next to one's ear and rotating it slowly about its axis of symmetry. One then hears alternately increases in loudness followed by diminution in loudness.

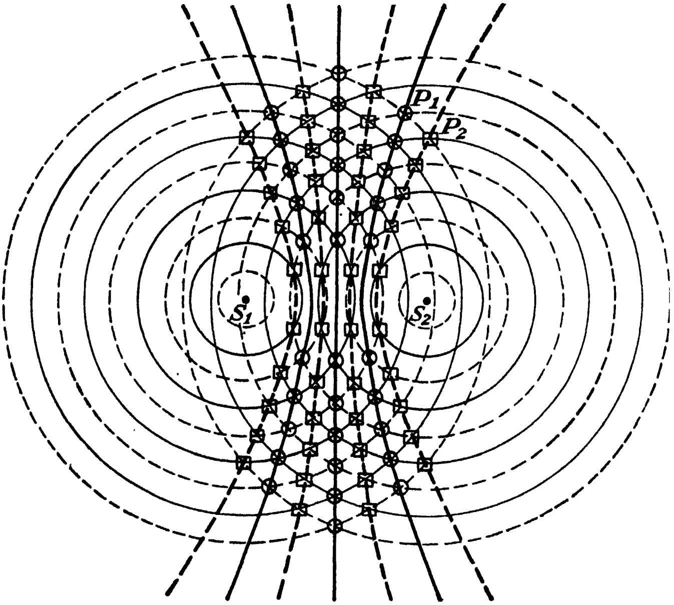


FIG. 29.2. Nodal (dotted curves) and antinodal (solid curves) lines in the interference pattern of two coherent point sources. (The solid circles represent crests, and the dotted ones troughs.)

Just as our ability to produce such a stationary interference pattern with sound demonstrates the wave nature of sound, our success in producing a similar stationary pattern of alternating zones of relative darkness and distinct illumination with the aid of two sources of light would be indeed very convincing evidence in favor of the wave theory of light. But everybody knows that no such pattern can be produced by illuminating a screen with two candles or two flashlights. This, however, does not disprove the wave theory of light! Suppose light is emitted from a source in the form of millions of short wave trains composed of a few wave lengths each. If we had two such sources, there would be, in general, no synchronization between the emitted wave trains. Sometimes the phase relation of the emitted waves would be such as to cause mutual reinforcement of the waves when they reach a given point on the screen. At other times they would cancel each other. Cancellation would be as frequent as reinforcement, and no interference pattern would be observed. **Only the use of two "coherent sources"**—*i.e.*, **of sources which emit waves with a constant phase relationship, would ensure a stable interference pattern.** Young conceived of an ingenious idea to secure two coherent sources of light. He illuminated a double slit (two closely spaced slits) by the same source of light. The two slits then acted like two sources vibrating in constant phase relation and produced a pattern of dark and bright fringes, which proved the wave nature of light and allowed the determination of the wave length of light.

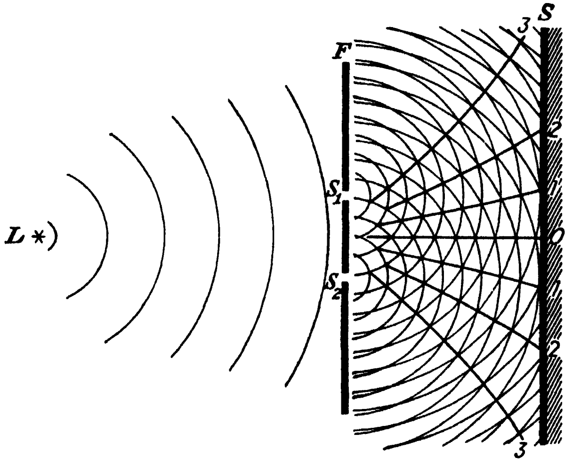


FIG. 29.3. Interference pattern obtained by using two slits as coherent sources of waves.

Young's Double-slit Experiment. The following experiment with water waves explains why Young's double slit is equivalent to two coherent sources of light waves (Fig. 29.3). L is a source of water waves. F is a barrier projecting out of the water with two openings S_1 and S_2, to which we shall refer as "slits." Since the two slits are equally distant from L, the wave which reaches them will agitate the water in the two openings in phase and with equal amplitude. According to Huygens' principle, the oscillating water particles in the two slits become sources of Huygens' wavelets. The two systems of circular waves issuing into the space to the right of the barrier F interfere with each other, forming nodal and antinodal regions similar to those indicated in Fig. 29.2.

Now let us replace L, the source of water waves in Fig. 29.3, by a light source of a definite color, and let us imagine ether waves to emanate from L. We shall expect now that a screen S placed in front of the double slit will not be uniformly illuminated by it. *Those parts of the screen which coincide with the nodal regions will be dark, and those coinciding with antinodes will be bright.* The antinodal points are connected by solid lines in Fig. 29.3. At the point where they coincide with the screen S, we find 0, 1, 2, 3, the bright fringes. Dark fringes (nodal regions) are located between the bright fringes. Young, who performed this experiment, actually detected a system of bright and dark fringes on the screen S. Figure 29.4 shows the appearance of such fringes which look straight as long as the length of the fringe considered is short as compared with the distance between the slits and the screen.

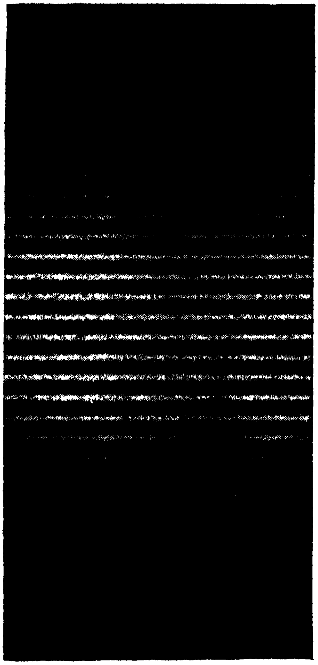


FIG. 29.4. Photograph of interference fringes obtained with a double slit (Young's experiment). (*From Jenkins and White, "Fundamentals of Optics," McGraw-Hill.*)

Determination of the Wave Length of Light

Figure 29.5 explains how it is possible to determine the wave length of light from the observed spacing of the fringes. Let us examine the illumination produced at various points of the screen S by the two coherent sources S_1 and S_2 (double slit), which we assume to be in phase. P is a typical point on the screen. Light coming from S_2 to P travels farther than light coming from S_1. We shall call the path difference Δ. When P coincides with point O, $\Delta = 0$. For the other positions of P an approximate formula can be derived, expressing Δ as a function of the distance x of P from the mid-point O, which is obtained as the intersection between the plane of the screen S with the perpendicular erected at the point c midway between the slits. This formula represents a good approximation as long as the angle α between the lines cP and cO is very small and the separation d of the slits is negligible as compared with the distance L of the slits from the screen.

Let us describe a circular arc with the radius $R = PS_1$ about the point

P. The point B is the intersection of this arc with the line PS_2. Hence $PS_1 = PB$, and the distance $S_2B = \Delta$ is the path difference referred to above. When the angle α is very small, the arc S_1B coincides very closely with the perpendicular dropped from the point S_1 upon the line PS_2. The angle S_2S_1B is very nearly equal to the angle $PcO = \alpha$, since $d \ll L$ so that the lines PS_1, PS_2, and Pc are nearly parallel to each other. Under these conditions we can assume with a reasonable degree of approximation that line $S_1B \perp Pc$ Hence, since $Oc \perp S_1S_2$, angle $S_2S_1B = PcO = \alpha$. We now have two similar triangles: (PcO) and (S_1S_2B). From triangle (PcO) it follows that

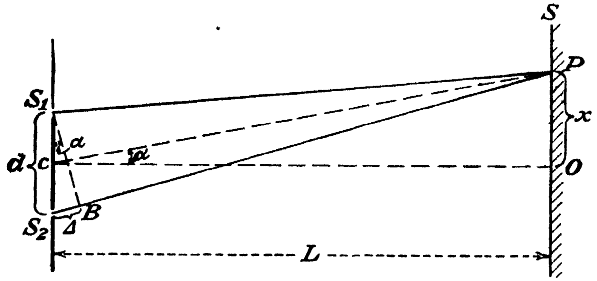


Fig. 29.5.

$$\tan \alpha = \frac{x}{L}$$

From triangle (S_1S_2B) follows

$$\sin \alpha = \frac{\Delta}{d}$$

If α is very small, $\sin \alpha \simeq \tan \alpha$, and we can write

$$\frac{\Delta}{d} = \frac{x}{L}$$

or

$$\boxed{x = \frac{L}{d}\Delta} \qquad (29.1)$$

If we now remember that light is a wave motion, we shall realize that, according to whether Δ is equal to an integral multiple of a wave length or an odd multiple of a *half* wave length, the waves will be either in phase or in phase opposition and will reinforce or cancel each other at P. We get thus as a **condition for dark fringes**

$$\Delta = n\frac{\lambda}{2} \qquad (n \text{ odd})$$

Substituting this value for Δ in equation (29.1) we get

$$\boxed{x = \frac{L}{d}\left(n \frac{\lambda}{2}\right)} \qquad (29.2a)$$

And for **bright fringes** we obtain, setting in equation (29.1) $\Delta = n\lambda$ (n = any integer), the condition

$$\boxed{x = \frac{L}{d}\,(n\lambda)} \qquad (29.2)$$

The following example of an actual measurement will give an idea of the order of magnitude of the physical quantities involved in such measurements. Using yellow glowing sodium vapor as the source of light and choosing $d = 0.59$ mm, one observes for the first dark fringe, $x = 1.0$ mm with $L = 2{,}000$ mm.

Setting $n = 1$ (for *first* dark fringe) in equation (29.2*a*) and solving for λ we get

$$\lambda = \frac{2x}{n}\frac{d}{L} = \frac{2 \times 1}{1} \times \frac{0.59}{2000} = 0.00059 \text{ mm}$$

for *the wave length of the yellow sodium light.*

It is customary to measure wave lengths in a special small unit of length: the **angstrom** unit (abbreviated as A).

$$\boxed{1\text{A} = 10^{-8} \text{ cm}}$$

Thus, the above value of λ is 5,900 A.

Measurements of λ for visible light reveal that *the range of visibility extends from about* 4000 *A* (*violet*) *to* 8,000 *A* (*red*). *It comprises only one octave. The wave length of light increases from the violet to the red end of the spectrum.*

If we use a white source instead of a monochromatic ("single-colored") one, we see a system of colored fringes. This can be understood as follows: White light is a mixture of light of different colors. Those places where the red component of white light is canceled by interference (*i.e.*, at the points at which the dark fringes of the red interference pattern should be located) will appear in the color complementary to red; they will appear blue-green. And similarly, cancellation of yellow light by interference (at different values of x from the case of red light) will give rise to blue fringes where monochromatic yellow light would have given rise to dark fringes.

The Diffraction Grating

The diffraction† patterns produced by a double slit are very faint and difficult to observe. A much brighter pattern can be produced by aligning a very large number of slits at regular intervals (**diffraction grating**). Such an arrangement could be produced, for instance, by painting parallel black lines on a glass plate. The clear interspace between the black lines would act like slits. In practice, one scratches the glass rather than painting it. The uninjured glass strips between the scratches act as parallel slits. It is possible to produce up to 30,000 scratches per inch.

If we look through such a diffraction grating at a monochromatic point or line source of light, we see several source images instead of one. And if we send a beam of a parallel monochromatic light through the grating, we observe several transmitted beams instead of one beam (Fig. 29.6).

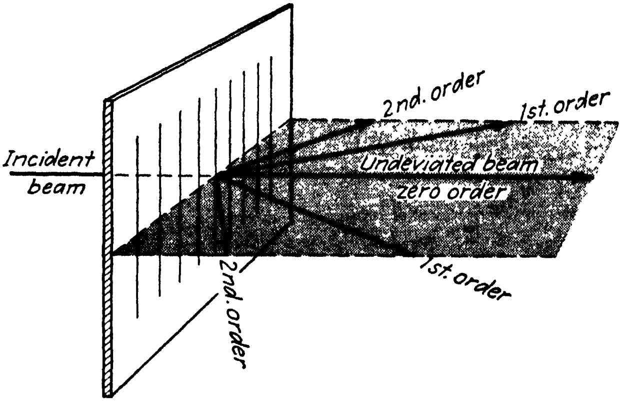


FIG. 29.6.

These two phenomena have a common explanation on the basis of Huygens' principle. Consider the cross section of the grating shown in Fig. 29.7*a*. A wave front *W* parallel to the plane of the grating, arriving from the left, excites ether vibrations in the openings of the slits, which thus become coherent sources of Huygens' wavelets. According to Huygens' principle, we obtain the resulting wave front on the right side of the grating by drawing an envelope (a common tangent) to the Huygens' wavelets *connecting points of equal phase with each other.* As we shall see presently, however, this can be done in several different ways.

Figure 29.7*a* shows the consecutive wave crests issuing from the slits (the troughs are omitted). We can, for instance, draw a tangent which passes through the *n*th crest of the wave emanating from slit 1, the *n*th crest of slit 2, etc. This will give us a line parallel to the grating surface, and, hence, also parallel to the original wave front. The wave progresses at right angles to the wave front, *i.e.*, in its original direction of propagation. The fronts of the transmitted wave are indicated by dashed lines (only two transmitted wave fronts are drawn in the figure). The fronts of the incident wave are indicated by solid lines.

---

† The term "diffraction" is applied to phenomena where bending of light into the region of geometrical shadows is observed.

It is, however, also possible to draw a common tangent to Huygens' wavelets in such a way that, instead of making it touch the nth crest of the waves originating from each slit, we make it tangent to the nth crest of slit 1, to the $(n + 1)$st crest of slit 2, $(n + 2)$nd crest of slit 3, etc. Two of such tangents are indicated by dashed lines in Fig. 29.7*b*. (For the sake of clarity of the figure, not all of the Huygens' wavelets have been drawn fully, and in Figs. 29.7*b* and *c*, the wave crests which lie ahead of the tangents have been omitted.) The wave progresses at right angles to the wave front, which, in this case, moves in a direction different from the direction of the original wave. It is clear for reasons of symmetry that another wave front could also have been constructed, deviated upward as much as the shown wave deviates downward. These two waves correspond to the two rays of Fig. 29.6 marked *first order.*

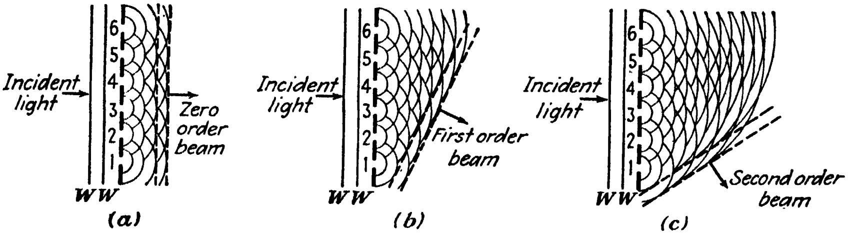


FIG. 29.7. Construction of the wave fronts for waves diffracted by a grating. (The Huygens' wavelets have been drawn only partially for greater clarity of construction.) *a.* Wave front of the zero-order beam. *b.* Wave front of the first-order beam. *c.* Wave front of the second-order beam. The slits of the grating are numbered.

Figure 29.7*c* shows another possible way of constructing a wave front, illustrating the so-called *diffraction of the second order.* We draw a common tangent to the following crests: nth crest of slit 1, $(n + 2)$nd of slit 2, $(n + 4)$th of slit 3, etc. We thus obtain wave fronts parallel to the dashed lines indicated in Fig. 29.7*c*. The direction of propagation of these second-order waves differs from the direction of the incoming wave much more than the direction of the first-order waves.

Similarly, under certain conditions (see page 524), still more tangents can be drawn corresponding to diffracted waves of the third, fourth order, etc. The diffracted waves of higher order diminish in intensity with increasing order.

From the direction of propagation of the diffracted waves, it is possible to determine the wave length of the diffracted light. Figure 29.8*a* explains how it can be done with the wave of the first order, and Fig. 29.8*b* with second-order waves. The lines marked W are the wave fronts, and the lines marked R are the rays, *i.e.*, perpendiculars to the wave fronts. Figure 29.8*a* is essentially the same as Fig. 29.7*b*.

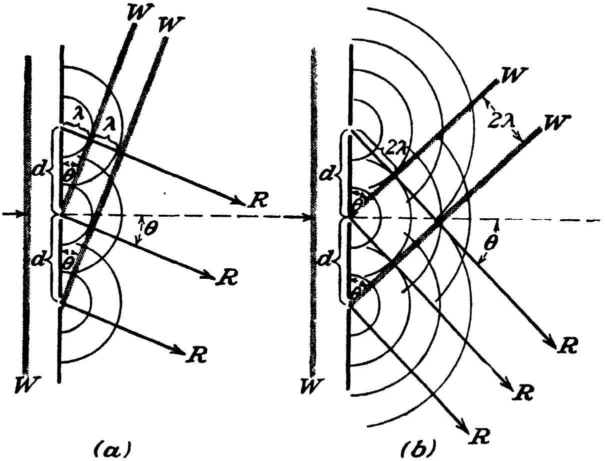


FIG. 29.8.

We see by inspection of Fig. 29.8*a*

$$\sin\theta = \frac{\lambda}{d}$$

If d is known, $\lambda = d\sin\theta$ can be determined by ascertaining the direction (angle θ) of the diffracted first-order wave (see Chap. 31, Grating Spectroscope).

In Fig. 29.8*b* the distance between two wave fronts is equal to 2λ.

From the triangle in which d is the hypotenuse and 2λ the side opposite the angle θ, we obtain

$$\sin\theta = \frac{2\lambda}{d}$$

or

$$2\lambda = d\sin\theta$$

In the case of a diffraction of the nth order we draw a diagram analogous to Fig. 29.8*b* making the distance between the wave fronts $n\lambda$ rather than 2λ. We can then write an equation similar to the preceding one:

$$\boxed{n\lambda = d\sin\theta} \tag{29.3}$$

This equation relates the angular deviation θ of the diffracted wave of the nth order to the wave length. The distance d between two neighboring slits (the so-called *grating constant*) is assumed to be known. Solving the equation (29.3) for n we see that the number of diffracted waves is limited:

$$n = \frac{d}{\lambda} \sin \theta$$

Since $\sin \theta$ cannot exceed the value of 1, *n cannot be any larger than d/λ. Thus, the highest possible order of diffraction is given by the ratio of grating constant over wave length.*

Diffraction by a Single Slit

The above theory of the diffraction grating was greatly simplified. It assumed that the slit is of negligible width as compared with the distance between successive slits and with the wave length of light. The practical gratings do not satisfy these assumptions perfectly, and the process of diffraction which they exhibit is more complicated owing to the fact that each individual slit of the grating gives a diffraction pattern of its own. The detailed theory shows, however, that, in the case of great line density, equation (29.3) represents a practically adequate formula for the direction of the diffracted waves.

Since the diffraction of light by a single slit is an important fundamental process, we shall consider it now in detail.

A superficial study of shadows led us to the conclusion that light is propagated in straight lines. According to this conclusion we expect the illuminated area of screen S to be exactly as wide as the slit if the slit in the wall W of Fig. 29.9 is illuminated by parallel light. The expected width of the illuminated area is obtained by drawing the dotted rays A and B through the edges of the slit. Actually we find, however, that the illuminated area is considerably wider if a very narrow slit is used (the slit width being equal to several wave lengths of the illuminating light). And in addition to the expected central illuminated area, we see several bright fringes whose brightness diminishes with their distance from the central fringe and which are separated from each other by dark fringes. The appearance of the illuminated screen S is shown in the photograph F of Fig. 29.9. The curve to the left of F shows the intensity of illumination in the region of alternating brightness and darkness.

The penetration of light into the geometrical shadow is referred to as **diffraction of light.** This observation cannot be accounted for in a simple fashion by the corpuscular theory of light. The early champions of the corpuscular theory tried to explain the observed bending of light around the edges of the slit by considering this process as refraction of light. They assumed that the ether in the vicinity of the edges of a material medium (such as the edges of the slit) was in a state of condensation, with the ether density increasing toward the surface of the medium and reaching a constant value in its interior. This "ether atmosphere" of varying density was supposed to refract the light toward the edges of the slit. But investigation showed that diffraction is independent of the index of refraction. This makes the above-mentioned explanation rather unconvincing. An additional difficulty of the corpuscular theory consists in its failure to predict the lateral maxima and minima of illumination.

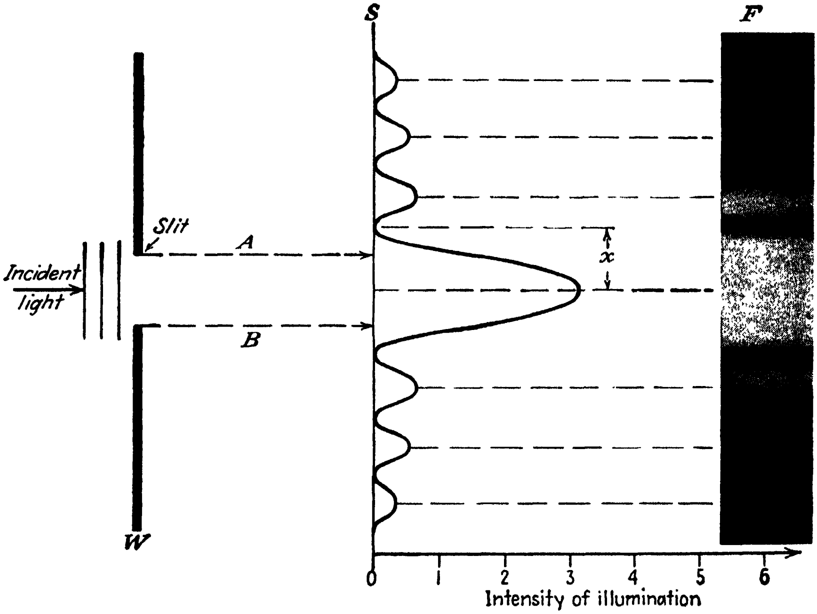


FIG. 29.9. Diffraction pattern obtained with a single slit. S, distribution of the light intensity; F, photograph of the diffraction pattern.

Unlike the corpuscular theory, the wave theory offers an elegant explanation of diffraction phenomena. Let us consider Fig. 29.10. The slit of width D is illuminated from the left by parallel light. The distance L of the screen S from the slit is very large compared with D ($L \gg D$). The ether particles in the plane of the slit can be considered as coherent sources of Huygens' wavelets. The illumination of every point of the screen can be considered as being due to the summation of light issuing from the individual vibrating ether particles lying in the plane of the slit. We shall show now that there are points on the screen at which the Huygens' wavelets emanating from the slit all destroy each other by interference. The points of the screen at which this happens are the points of the dark fringe regions. At other points the cancellation is not so complete, and we get regions of relative brightness. A maximum of brightness is located between two brightness minima. *We shall limit ourselves here to the computation of the distance x (see Figs.* 29.9 *and* 29.10) *of the first minimum from the center.*

The location of the first minimum can be determined as follows (Fig. 29.10): Divide the slit into two halves. Designate vibrating ether particles in the upper area of the slit opening by *italic* letters (a, b, c, . . .) and those of the lower area by Greek letters (α, β, γ, . . .). Draw rays from corresponding pairs of points (such as a and α, b and β, etc.) toward

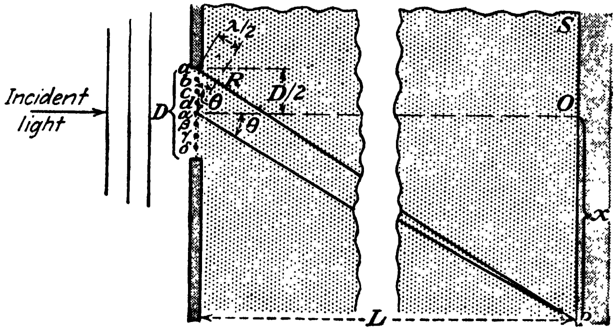


FIG. 29.10.

a point P on the screen S. If the distance L between the screen and the slit is very much larger than the slit width D, *these rays will be nearly parallel.*

Under these conditions the perpendicular dropped from the point α upon the line aP coincides very nearly with the circular arc αR obtained by describing a circle of radius $r = \alpha P \simeq RP$. The distance aR is thus practically the amount by which the path aP exceeds the distance αP. The distance which the light from a has to travel to P is thus larger than the distance which the light from α has to travel to P. If the path difference is $n(\lambda/2)$ (where n is an odd integer), the waves from a and α will arrive 180° out of phase at P and will annul each other. The same construction can be applied to all the other pairs of "corresponding points" (b and β, c and γ, etc.). This construction shows that, if light waves coming from a and α arrive at P in phase opposition, the waves arriving from the pairs of sources (b and β, c and γ, d and δ) will also be in phase opposition and hence will cancel each other at P. Thus, we see that there will be a certain angle θ for which the path difference between any two "corresponding rays" (like a and α) will be equal to $\lambda/2$. Points of the screen illuminated by such rays will be dark. This illustrates the formation of the first dark fringe.

By measuring the distance x of the first dark fringe from the center of symmetry O of the fringe pattern, the wave length of light can be computed from the following relations:

$$\tan\theta = \frac{x}{L} \quad \text{(from the triangle } \alpha OP\text{)} \qquad (29.4a)$$

and from the small triangle $a\alpha R$ we get

$$\boxed{\sin\theta = \frac{\lambda/2}{D/2} = \frac{\lambda}{D}} \qquad (29.4b)$$

for small angles, $\sin\theta \cong \tan\theta$; hence (approximately),

$$\frac{\lambda}{D} = \frac{x}{L}$$

and

$$\boxed{\lambda = \frac{D}{L}x} \qquad (29.5)$$

This equation allows us to determine the wave length of light by measuring the distance $2x$ between the two first minima, the distance L of the slit from the screen and the width of the slit.

Let us now consider equation (29.4*b*): $\sin\theta = \lambda/D$. Since θ determines the location of the first minimum, it can be considered as an indication of the extent of the central illuminated area, which is terminated on both sides by the first minima. We see that (at constant λ), as we make the slit width D smaller, θ grows larger, *i.e.*, the central illuminated area becomes wider. $\lambda/D = 1$ is evidently the maximum possible value which this ratio can assume. This occurs when $\sin\theta = 1$; that is, $\theta = 90°$. This means that, when D becomes equal to the wave length of light ($\lambda/D = 1$), we get $\theta = 90°$; that is, the entire screen S is illuminated by the slit. On the other hand, for a given value of the slit width D, we see from $\sin\theta = \lambda/D$ that the central bright area becomes wider for greater wave lengths of light.

The bending of light into the region of the "geometric shadow" eliminates the argument of the proponents of the corpuscular theory that light cannot be a wave motion because it does not bend around corners as sound does. The bending of light is so much smaller than the bending of sound because of the smallness of λ for light as compared with the wave lengths of audible sound.

The above considerations not only apply to diffraction of light waves but hold equally well for waves of any kind.

A circular opening gives a diffraction pattern similar to that of a slit. In this case we observe a central bright circle surrounded by a series of concentric rings of alternating brightness (Fig. 29.11). It can be shown that the formula for the width of the central bright circle is given by an expression similar to equation (29.4*b*) for the width of the central bright fringe in the case of a slit. The two expressions differ by a factor of 1.22, which appears in the formula that applies to the circular opening of diameter D:

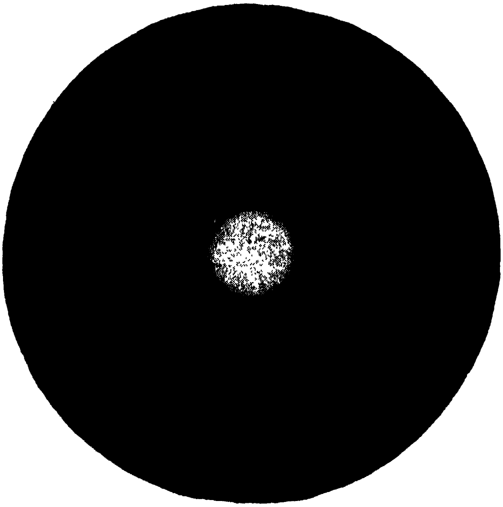


Fig. 29.11. Diffraction pattern obtained with a circular hole. (*From Monk, "Light: Principles and Experiments," McGraw-Hill.*)

$$\boxed{\sin \theta = 1.22 \frac{\lambda}{D}} \tag{29.6}$$

where $\sin \theta \simeq x/L$. Hence $x/L = 1.22\lambda/D$ (29.6*a*)

The significance of the angle θ in this equation becomes clear if we imagine Fig. 29.10 rotated about the central axis $\alpha 0$. This makes D a circular opening, and x becomes the radius of the circle described by the point P. If the waves arrive in phase opposition at P, x is the radius of the first dark ring and hence also of the inner bright circle. θ is the angle which the radius x of the first dark circle subtends at the center of the circular aperture D.

Interference Due to Multiple Reflection of Light from Thin Films

, We shall now turn to phenomena which at first glance seem to be unrelated to the process described above but, however, turn out to be caused by interference of light in a similar way to the color effects which we observe when viewing a distant light source through a diffraction grating.

We frequently notice a pretty display of colors on a rainy day at places where a fine film of oil has been deposited on the wet pavement. A similar effect can be obtained by heating a piece of iron. After it cools, its oxidized surface shows a variety of colors. When we place a thin microscope slide cover on a glass plate and press it against the plate with a sharp point, we see colored loops around the point at which the pressure is applied. The lively changing colors of thin soap bubbles are well known to all (Fig. 29.12). What is the cause of this display of color by originally colorless materials? In our search for the answer to this question, let us begin by considering the following experiment which has been performed by Newton.

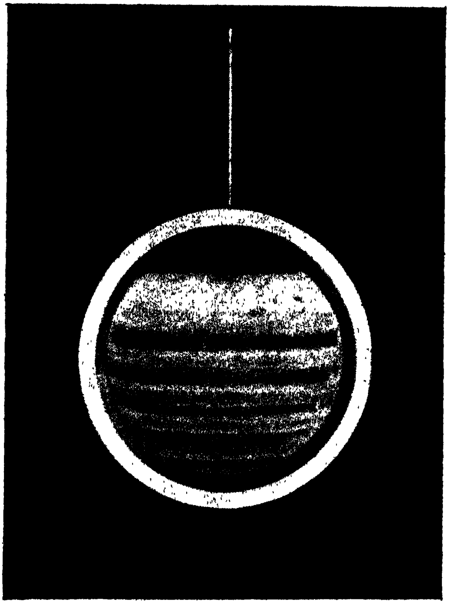
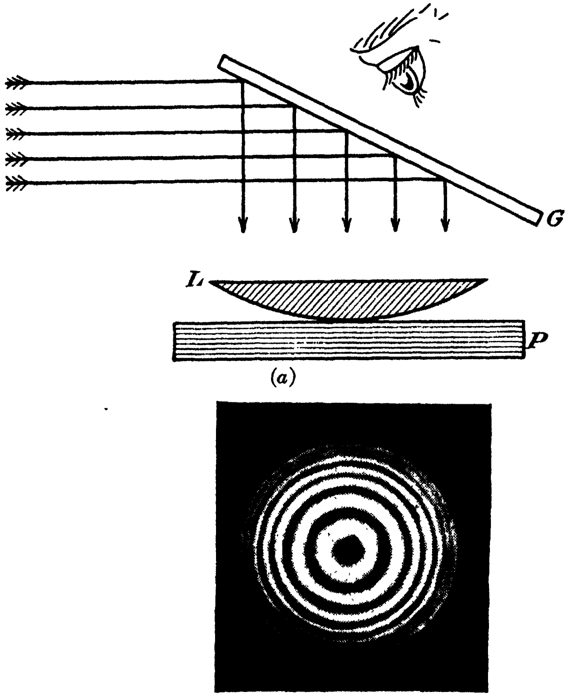


FIG. 29.12. Colored interference pattern obtained with a vertically mounted soap film. (*Courtesy of R. A. Regalbuto.*)

A glass lens L of small curvature is placed on a flat glass plate P (Fig. 29.13a). This setup is illuminated by parallel *white* light which is issuing from a source on the left and is reflected downward by the glass plate G. An observer looking from the top then sees a series of concentric *colored* rings (**Newton's rings**). If, instead of using white light for illumination, we use monochromatic light, we see alternately dark and bright rings (Fig. 29.13*b*). This picture reminds us of the alternation of illumination in the diffraction pattern produced by light passing through a circular hole. The formation of that pattern was explained as being due to interference of light waves. The similarity of the two phenomena suggests that Newton's rings may also be an interference effect. This suspicion is strengthened when we observe that the radii of Newton's rings increase if the color of illuminating light changes from blue to red.



(*a*)

(*b*)

FIG. 29.13*a*. Observation of Newton's rings. *b*. Newton's rings.

A similar widening of the pattern is observed when diffraction patterns of a slit and circular opening are observed with blue and with red light.

But for interference to take place we need at least two coherent sources of light. Where do we have two such sources in Newton's experiment? Light which passes downward from the plate *G* is partially reflected at every glass surface which it encounters. The air space between the lens and the flat plate has the cross section of a thin curved wedge whose thickness increases outward. The curved surface of the lens and the upper surface of the glass plate are the boundaries of this air wedge. Some light will be reflected upward from the top boundary and some from the bottom boundary of this wedge. *Thus we shall have two coherent* *reflected waves which can interfere with each other.* In order to understand how the interference between these two waves leads to the ring pattern, we shall take two intermediate steps by considering interference patterns obtained with (1) a thin plate and (2) a thin linear wedge.

Figure 29.14 shows a plane parallel glass plate *G* which is illuminated obliquely by parallel light. Only one of the incident light rays is pursued in its course. The *thin* slab of glass appears *uniformly colored* when illuminated by parallel *white* light. This can be interpreted as follows: Every incoming ray is split into two rays: ray 1, which is reflected from the top surface, and ray 2, which makes the detour *Oab* before emerging into the air after a reflection at the bottom surface. As a result the light vibrations in ray 2 will lag in phase behind the vibrations in ray 1. If path *Oab* is of such a length that light vibrations in rays 1 and 2 are 180° out of phase for a certain color, the two rays will cancel each other in their effect upon the eye of the observer. For this particular value of the angle *i* and for light of that particular wave length for which this cancellation condition holds, the plate will appear dark in reflected light. If the illuminating light is white, light of the wave length for which the cancellation condition holds will be eliminated by interference and the plate will appear in the complementary color.

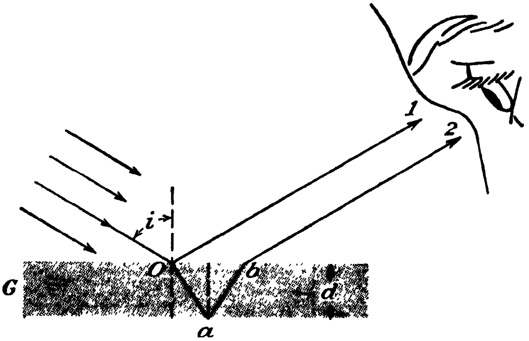


FIG. 29.14. Reflection of incident light from both surfaces of a glass slab.

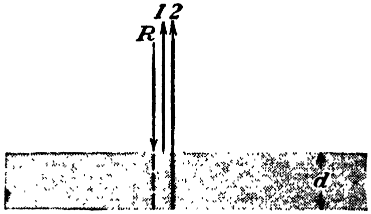


FIG. 29.15.

The situation is particularly simple in the case of perpendicular incidence of light. Consider a schematic diagram for the case of perpendicular incidence of ray *R* (Fig. 29.15). Actually ray *R* and the two reflected rays 1 and 2 are to be pictured as coinciding, but in this diagram they have been separated for clarity. Light which is reflected from the back surface (ray 2) traverses twice the thickness of the plate and thus makes a detour $\Delta' = 2d$ as compared with the path of the light which is reflected from the front surface (ray 1). We should expect rays 1 and 2 to weaken or cancel each other by interference if $2d = \lambda^*/2$† (*i.e.*, if the path difference is one-half of the wave length in glass), since then the two rays 1 and 2 will be in phase opposition. Thus we should expect not to see any reflected light for $d = \lambda^*/4$, since then $\Delta' = \lambda^*/2$. Complete cancellation of reflected light could occur only if the intensities of rays 1 and 2 are equal. If they are not (which is generally the case), we should obtain only partial cancellation.

Actually, this is *not* what we observe when we study films of different thickness. When $d = \lambda^*/4$, we observe the *maximum* intensity of reflected light. However, when $d \approx 0$ (d negligible as compared with the wave length of light) we get *minimum* intensity of reflected light. The second result is also contrary to our expectation, since for $\Delta = 0$ both reflected waves should be practically in phase and should intensify each other.

This result can, however, be explained if we assume a phase reversal of the vibrations of light when the latter is reflected from one of the two surfaces, the front or the back surface.

In Chap. 27, in connection with the discussion of reflection of waves in ropes (see Figs. 27.12 and 27.14), we saw that the **reflection of a wave from a denser medium** (heavier rope) **is accompanied by phase reversal whereas reflection from a less dense medium** (lighter rope) **is not.** This analogy led Young to the conclusion that *the phase of the vibrations of light is reversed when they are reflected back into vacuum at an interface vacuum glass.* This assumption of phase reversal enabled Young to explain why the center of the pattern of Newton's rings is black in reflected light *for any color* in spite of the fact that the thickness of the air wedge is nearly zero at that point. But Young had no evidence except the above-mentioned analogy for the assumption that the phase reversal takes place when light impinges upon an "optically denser" medium (*i.e.*, one of the higher refractive index) rather than when it falls upon

---

† λ^*, the wave length *in glass*, is related to the wave length *in a vacuum* as follows: $\lambda^* = \lambda/n$, where n is the refractive index of the glass. This follows from the relation $v = f\lambda^*$, where v is the velocity of light in the given medium. For the same light in a vacuum: $c = f\lambda$.

hence
$$\frac{c}{v} = \frac{f\lambda}{f\lambda^*}$$

and since $c/v = n$ by definition, it follows that

$$\lambda^* = \frac{\lambda}{n}$$

an "optically less dense" medium. A more convincing support for Young's hypothesis was presented much later by Wiener.

Wiener reflected a monochromatic light wave incident perpendicularly on a metal mirror and obtained a pattern of standing light waves owing to interference between the incoming and reflected light. The nodal (dotted lines) and antinodal (solid lines) surfaces (Fig. 29.16) formed a series of planes parallel to the mirror surface (*M*). He was able to distinguish the light nodes from the antinodes by placing an exceedingly thin photographic film *P* obliquely in front of the mirror. The film was blackened at regular intervals (see Fig. 29.16)† which were labeled antinodes of ether vibrations by analogy to mechanical waves. Wiener found no blackening at the mirror surface and thus concluded that there was a nodal region at the mirror similar to the node at the fixed end of the string in Fig. 27.10. As we saw before, a node at the reflecting surface indicates phase reversal on reflection. Thus Wiener's experiment vindicated Young's assumptions.

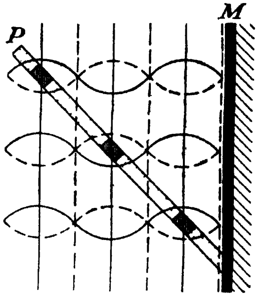


FIG. 29.16. Wiener's experiment with standing light waves.

We can now determine the conditions under which the plate shown in Fig. 29.15 will appear dark or bright in reflected light. The phase reversal of ray 1 has the same effect as an additional detour of ray 2 by $\lambda^*/2$ in the glass. If we now add this fictitious detour of ray 2 to its actual detour of $\lambda^*/2$ (which we obtained in the case of $d = \lambda^*/4$), we see that both rays emerge into the air with a phase difference of $2\lambda^*/2$ or λ^*, which means that they are in phase.

On the other hand, for $d \approx 0$, the actual detour of ray 2 is zero. There is, however, a phase difference of $\lambda^*/2$ between the two rays due to phase reversal on reflection of the ray 1.

In general, there will be destructive interference of the rays reflected from the front and back surface if the "effective detour" Δ (including the fictitious detour due to phase reversal on reflection) is an odd multiple of $\lambda^*/2$:

$$\boxed{\Delta = \underset{\substack{\uparrow\\ \text{(odd}}}{N\frac{\lambda^*}{2}} = \underset{\substack{\uparrow\\ \text{actual}\\ \text{detour}}}{2d} + \underset{\substack{\uparrow\\ \text{fictitious}\\ \text{detour)}}}{\frac{\lambda^*}{2}}} \qquad (29.7)$$

---

† The incoming waves (solid curve) and the reflected light waves (dotted curve) are shown in the figure as transverse waves to illustrate the formation of the nodal and antinodal surfaces.

Setting $N = 1$ in equation (29.7) we see that the first destructive interference occurs for $d \approx 0$. The second one for $N = 3$ occurs when $2d = (3 - 1)\lambda^*/2$ or $d = (3 - 1)\lambda^*/4 = \lambda^*/2$. The third one for $N = 5$ occurs when $d = (5 - 1)\lambda^*/4 = \lambda^*$, etc. Thus, if the thickness of the plate of Fig. 29.15 were to increase continually as we look at it from the top, it would appear alternately dark and bright in reflected monochromatic light. If white light were used, it would appear blue-green when its thickness was such as to make it appear dark in red light.

The change in appearance of a glass plate with increasing thickness can be demonstrated by looking at a glass wedge. If the boundaries of the glass slab of Fig. 29.15 are not exactly parallel, the slab can be considered as a wedge. If the angle between the surfaces of the wedge is very small, alternate dark and bright fringes become visible when the plate is illuminated by parallel monochromatic light as shown in Fig. 29.17. Dark fringes appear where the thickness of the wedge is an even multiple of $\lambda^*/4$, and bright ones appear where the thickness is an odd multiple of $\lambda^*/4$. Where the thickness is the smallest even multiple of $\lambda^*/4$, namely, $d = 2(\lambda^*/4) = \lambda^*/2$, the actual detour is $2d = \lambda^*$. Hence, the two reflected waves (corresponding to rays 1 and 2 of Fig. 29.15) should be in phase. But they are not because of the phase reversal on reflection from the "optically denser" medium. The two reflected waves are in phase opposition, and hence, the wedge appears *dark* in reflected light at points where $d = \lambda^*/2$. It is clear that, if we increase d by $\lambda^*/2 = 2\lambda^*/4$, the detour will be increased by $2d = \lambda^*$, so that the phase relation between the two reflected rays will remain unchanged.

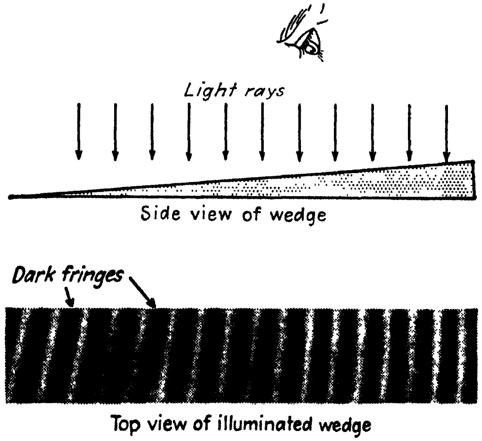


FIG. 29.17. The photographed portion is taken from Monk, "Light: Principles and Experiments," McGraw-Hill.

In other words the wedge will appear dark at points where d is an integral multiple of $\lambda^*/2$ or, which is the same, an even multiple of $\lambda^*/4$.

On the other hand at points where $d = \lambda^*/4$, the detour is $2d = \lambda^*/2$ and the "effective detour" is $\Delta = (\lambda^*/2) + (\lambda^*/2) = \lambda^*$. The two reflected waves are in phase, and the wedge appears bright in reflected light at these points if illuminated as shown in Fig. 29.17. Since increasing d by $\lambda^*/2$ does not change the phase difference between the two reflected rays, as was shown above, the wedge will appear bright at points where the plate thickness is

$$\begin{aligned} &\lambda^*/4 \\ &(\lambda^*/4) + (\lambda^*/2) = 3\lambda^*/4 \\ &(\lambda^*/4) + (2\lambda^*/2) = 5\lambda^*/4, \text{ etc.} \end{aligned}$$

Thus, in general, bright fringes will appear whenever d is an odd multiple of $\lambda^*/4$.

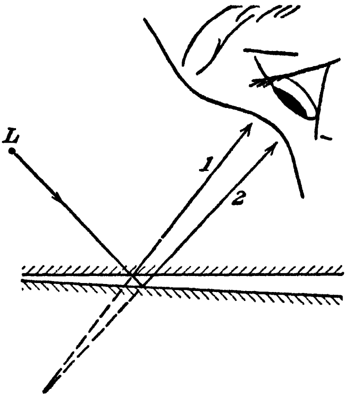


FIG. 29.18.

If white light is used for illumination, the wedge will show colored fringes. For instance, at those points where the red light waves are canceled by interference, the wedge will appear in the complementary color. Figure 29.12 is an illustration of this effect. The soap film assumes a wedge shape owing to gravity when placed vertically, the bottom being thicker than the top. In Fig. 29.12 the top is so thin that all colors are canceled there (since the fictitious detour is the same for all). The rest of the film appears in changing colors as its thickness changes. Figure 29.18 shows the two interfering rays 1 and 2 reflected from the two surfaces of a wedge entering the pupil of the observer's eye.

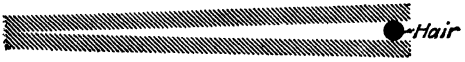


FIG. 29.19.

A similar observation can be made with an air wedge. Such an air wedge is obtained, for instance, by placing a human hair between two glass plates as shown in cross section in Fig. 29.19. If the surfaces of two glass plates which are in contact are not perfectly plane but, rather, somewhat "hilly," one sees an irregular pattern of interference fringes as shown in Fig. 29.20. This is a very sensitive test for optical smoothness of a glass surface.

We are now in a position to interpret Newton's rings shown in Fig. 29.13. The rings are the interference pattern formed by reflection of light from the two glass boundaries of the air space of wedge-shaped cross section separating the lower surface of the lens from the top of the glass plate. The center appears black because the air gap between the lens and plate is of zero thickness at the point of contact. The ring-shaped pattern of dark and bright fringes is explained in the same fashion as the linear interference fringes observed with a solid wedge or air wedge.

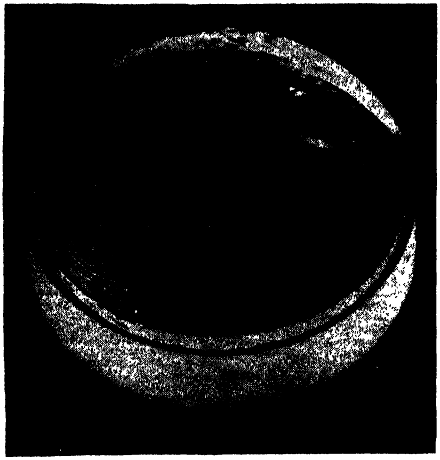


FIG. 29.20. Interference pattern obtained with an air gap of uneven thickness. (*Courtesy of Bausch and Lomb Optical Company.*)

Interference phenomena observed with thin reflecting films have found recently an important practical application in the preparation of the so-called *nonreflecting glass*. The reflection of light from the surfaces of lenses and other optical elements in optical instruments, such as the microscope, camera, telescope, and spectroscope, causes a considerable amount of stray light and loss of useful light, thus rendering the instrument less effective. The reflection can be made negligible by coating the glass surface with a very thin layer of a material which has an index of refraction smaller than that of glass. Then light reflected from both surfaces of the film will be reflected back into an optically less dense medium and will suffer a phase reversal *in both cases.* If the film thickness d is $= \lambda^*/4$, the normally incident light will traverse the distance $2d = \lambda^*/2$ in the film before it is united with the light reflected from the front surface. Hence, *both light waves are out of phase and cancel each other* if they are of equal intensity. The equality of their intensity can be achieved by a proper choice of the index of refraction of the film material. Thus the coated surface is rendered nonreflecting.

Obviously this condition can be achieved only for a definite wave length of light. One chooses the yellow-green region, which is the region of maximum sensitivity of the eye. The light losses due to reflection are thus reduced from about 5 per cent per reflecting surface to a fraction of 1 per cent.

We have seen in this chapter how simply a variety of intricate phenomena can be explained by the assumption that light is a wave motion in a hypothetical medium, the ether. An attempt to explain these phenomena on the basis of the corpuscular theory would require a series of assumptions about the properties of the corpuscles which would make the resulting theory more complicated than the wave theory. We should have to assume that corpuscles undergo some periodic changes in flight with frequencies which vary from color to color. We should have to assume, for instance, that in a vacuum the velocity of these different corpuscles is independent of color and of the temperature of the glowing body from which they are emitted. These assumptions appear very artificial in a corpuscular theory but are quite natural in a wave theory, since we know that the velocity of propagation of waves in a medium is nearly independent of the intensity of vibration of the source and can be essentially independent of its frequency over a wide range. If we adopt from here on the wave theory of light, it is not because we are unable to explain all the observed phenomena by assuming light to consist of corpuscles but because we can get along with a smaller number of assumptions in formulating a wave theory of light and can interpret and predict phenomena more simply.

QUESTIONS AND PROBLEMS

1. Describe a simple setup for the determination of the wave length of light. Suggest as many different methods as you can.

2. Draw the pattern of standing waves produced when light is reflected from a mirror at an angle of incidence of 45°.

3. What would you see on the screen if you were to illuminate the double slit simultaneously with a red and green source in Young's experiment?

4. Can you separate white light into its component colors by sending it through a diffraction grating? In what way will the result be different from the experiment in which the white beam is sent through a prism?

5. It can be shown that light penetrates into the geometrical shadow of a hair which is held in front of a point source of light. How do you explain it?

6. How would Newton's rings appear if two slightly different wave lengths of light were used for illumination?

7. Try to explain Newton's rings on the basis of the corpuscular theory.

8. Does the successful explanation of the interference phenomenon on the basis of the wave theory rule out a successful corpuscular theory of light?

9. Is it conceivable that the wave length of green light could ever be longer than the wave length of red light? (Discuss it.)

10. What will be the color of reflected light when a celluloid film 1.5×10^{-7} cm thick is illuminated at perpendicular incidence (*a*) by white light? (*b*) by red light?

11. How thick must a film of water be in order to appear green in reflected light (at normal incidence of white light)?

12. If a film of oil appears yellow in reflected light, what will be the color seen in transmitted light? (Explain why the transmitted light is colored.)

13. How do we infer the existence of the phase shift which occurs when light is reflected?

14. How do we know at which of the boundaries the phase shift occurs?

15. A plane parallel glass plate 0.0100 mm thick is illuminated by vertical incidence of white light. What is the color of the reflected light? (Use $n = 1.50$.)

16. A plane parallel glass slab of refractive index $n = 1.40$ fails to reflect light (appears dark) when illuminated by perpendicularly incident parallel light of wave length $\lambda = 4{,}000$ A. Can we infer from this how thick the plate is? (Make a quantitative statement.)

17. A human hair H is placed between two glass plates as indicated in Fig. 29.19. When the upper plate is illuminated by perpendicularly incident parallel light of $\lambda = 5{,}460$ A, 70 bright fringes appear across the plate (see lower part of Fig. 29.17). What is the approximate diameter of the hair?

18. Why can't you produce interference fringes by illuminating an area by two candles?

★19. How could you use two mirrors to measure the wave length of light?

20. If you send a beam of parallel yellow light rays ($\lambda = 6{,}000$ A) through a diffraction grating of 5,000 lines per centimeter, how many beams of light will be detected after passage of the original beam through the grating?

21. What do you see when you look through a diffraction grating at a glowing metal filament? Explain the observed phenomenon.

22. The double slit of Fig. 29.5 is illuminated by light of $\lambda = 5{,}000$ A. Assuming $L = 2$ m and $d = 0.5$ mm, determine the deviation of the second dark fringe from the center.

23. Solve Prob. 22 assuming the apparatus to be submerged under water.

24. A beam of light is sent through a diffraction grating of 6,000 lines per centimeter. The diffracted beam of the second order forms an angle of 30° with the central beam of light. What is the wave length of the light used?

25. Parallel light ($\lambda = 5{,}000$ A) is sent through a slit of the width $d = 0.1$ mm. As a result a bright line is seen opposite the slit on a screen 5 m away. How wide is this line? (Give the distance between the two first dark fringes.)

26. Solve Prob. 25 replacing the slit by a circular hole of 0.1 mm diameter.

CHAPTER 30

FORMATION OF OPTICAL IMAGES

a. What is an image?

b. What is a (1) virtual, (2) real image?

c. How is the image formed by reflection from a plane mirror?

d. How is light focused by a parabolic mirror?

e. How is light focused by an elliptical mirror?

f. What is the advantage of a spherical mirror over a parabolic and an elliptical mirror?

g. What is meant by "reversibility of the light path"?

h. Explain image formation by a spherical mirror.

i. When do we get a real image (1) with a convex spherical mirror? (2) with a diverging lens?

j. When do we get a magnified image with a (1) concave mirror? (2) convex mirror? (3) converging lens? (4) diverging lens?

k. When do we get a virtual image with a (1) concave mirror? (2) convex mirror? (3) converging lens? (4) diverging lens?

l. When do we get an upright image with a (1) concave mirror? (2) convex mirror? (3) converging lens? (4) diverging lens?

m. What is the relation between magnification and the positions of object and image?

n. Explain the action of a lens on the basis of the wave theory of light.

o. What factors determine the "speed" of a camera?

p. How do you explain the corrective action of eyeglasses?

q. How do you explain the action of a magnifying glass? Must the image be formed at $q = -25$ cm?

r. How do you explain the action of a compound microscope? What measures can be taken to reach the maximum possible magnification?

s. Explain the magnifying action of Kepler's telescope.

t. Why are astronomical telescopes usually very long and of large diameter?

u. What determines the limit of resolution achievable with an optical instrument?

v. What is meant by resolving power?

w. How can resolution be improved?

MIRRORS AND LENSES

One of the important objectives of physics is to aid our senses, *i.e.*, to correct for their imperfections and to extend the range of their perception. Thus, for instance, the telephone and microphone enable us to hear sounds produced at very great distance. The oscilloscope enables us to *see* high- and low-frequency sounds which we are unable to hear, and the thermometer helps us detect temperature changes which we could not feel without the aid of instruments.

Our sense of vision is unquestionably our most important link with our environment. It is responsible for the acquisition and extension of our knowledge to a greater extent than any other sense. The reasons for our desire to "sharpen" this sense by utilization of physical laws are quite obvious. We should like to see luminous objects too faint to be seen by the unaided eye, such as distant stars and island universes; we should like to see objects too small to be seen without optical aid, such as bacteria; we want to know if there are radiations similar to visible light to which our eye is insensitive; we want to correct the vision of people whose eyes are defective; and finally, the inability of our eye to analyze colored light into its component colors requires special instrumentation to accomplish this task. Devices serving any of these purposes are referred to as *optical instruments.* The first part of this chapter is devoted to the elements of which optical instruments may consist, namely mirrors and lenses.

The Plane Mirror

We shall begin with the simplest optical element, an instrument which enables us to see our own eyes, the plane mirror. The use of a plane mirror is to change the direction of propagation of light waves without distorting the shape of the wave front. Figure 26.14 shows how a plane wave front AB moving southeast is reflected and becomes a plane wave front DE traveling northeast. In the consideration of optical instruments, it is more convenient to consider *light rays* rather than light waves. Although the light rays can be taken to represent the path of light corpuscles, the ray representation does not necessarily commit us to the corpuscular theory of light. We can imagine obstacles placed in the path of the wave front which (obstacles) are large as compared with the wave length of light. These obstacles will cast shadows which will mark the path of our rays. $W_1 - W_6$ represent wave fronts in Fig. 30.1; A, B, C, D, represent the obstacles and $Aa\alpha$, $Bb\beta$, . . . , $Dd\delta$ the "rays," *i.e.*, shadows. We could just as well imagine W_1 to be a screen and A, B, C, D, openings in the screen. Then our rays $Aa\alpha - Dd\delta$ would be thin beams of light.

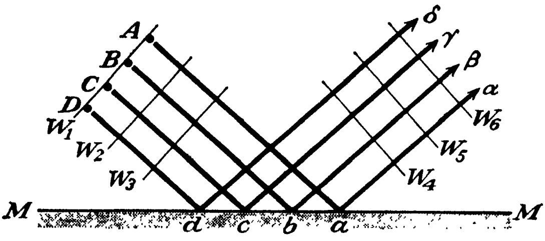


FIG. 30.1.

The light rays and the wave fronts form a system of mutually perpendicular lines analogous to the lines of force and the equipotential lines.

Suppose that, instead of considering reflection of plane waves from a plane mirror, we consider reflection of spherical waves emanating from a point source of light. Figure 30.2 shows an analogous experiment performed with a water wave model. Water waves emanate from O (solid circles) and are reflected when they reach the fence MM (mirror). The reflected waves have the shape of concentric circles whose common center of curvature lies on the opposite side of the mirror at I.† The location of this center of curvature I, which is called the **image point** of the point *object* O, can be accomplished by drawing normals (N_1 and N_2) to the reflected wave front at any two points. I is located at the point of intersection of the normals.

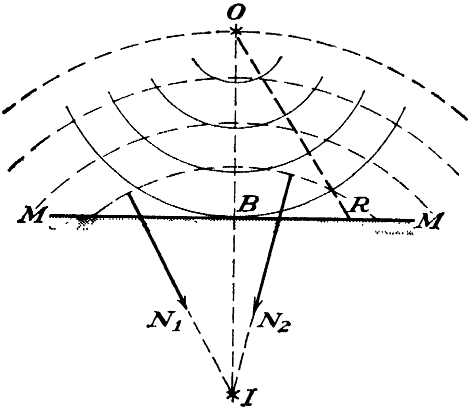


FIG. 30.2. Reflection of spherical waves by a plane mirror.

Formation of a Virtual Image. There is an easier method of locating the image point I. It can be found by considering the incident rays, *i.e.*, radii drawn from O (at right angles to the wave fronts), such as R in Fig. 30.2. The image point I is then found by locating the intersection of the backward prolongation of the reflected rays. Figure 30.3 illustrates how the law of reflection applied to the incident rays locates the image I. The incident rays are drawn in solid lines, the reflected rays are dashed, and the backward prolongations of rays are dotted.

The incident ray OB forms the angle $\theta' = 0$ with the normal. According to the law of reflection the angle of reflection is equal to the angle of incidence; *i.e.*, the reflected ray (BR) coincides with the incoming ray OB.

---

† The shape of the reflected wave front can be determined by the usual method of drawing the envelope of the Huygens' wavelets which originate at the surface of the mirror at the moment when the incident wave strikes the mirror.

OC is supposed to represent an arbitrary incident ray. The reflected ray CD is drawn according to the law of reflection. (Note equal angles θ on either side of N.) The reflected rays BR and CD are the two desired normals of the reflected spherical wave front. Hence, we find the center of curvature of the reflected wave front (*i.e.*, the image point I) by prolonging OB and DC backward until they intersect.

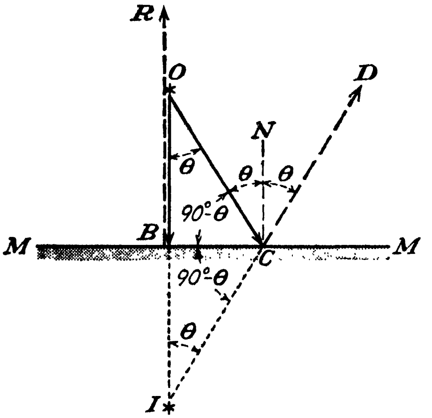


FIG. 30.3. Reflection of light rays from a plane mirror.

This consideration applies to all kinds of waves, including light waves. Our visual sense performs subconsciously the task of prolonging backward the rays which enter our eyes and of judging the source of light to be at the intersection of the backward prolongation of the rays: *the image obtained by intersection of prolonged rays rather than by actual intersection of rays is called a* **virtual image.** I is a virtual image of O in Fig. 30.3.

It is easy to prove that the virtual image lies on the prolongation of the normal OB and that it is as far behind the mirror as the object is in front of it. We have merely to prove the congruence of the triangles BCO and BCI. In Fig. 30.3 we see that they have one side in common (namely BC), are both rectangular, and that angle $BCI = 90 - \theta =$ angle OCB, because angle $DCI = 180°$ and angle $BCI = 180° - [(90 - \theta) + 2\theta]$. Hence, the two triangles are congruent, and consequently, $OB = IB$. Since OC can be any arbitrary ray issuing from point O, this result means that the backward prolongation of *all* reflected light rays intersect at a common image point I, which is as far behind the mirror as the object point O is in front of it.

This suggests a simple construction of the image point I. Drop the perpendicular from the object point O upon the mirror [or the prolongation of the mirror (see page 543)], double the object-mirror distance, and you arrive at the image point I. Now a luminous object can be considered as being composed of a multitude of luminous object points, and this construction can be carried out for each individual object point. Practically, the location of only a few representative points is sufficient for the construction of the image. Figure 30.4 shows, for example, the construction of the image of an arrow. The necessity of prolonging the line MM beyond the mirror surface to find the image point I_1 may suggest at first glance that point I_1 may not be visible, since the object point O_1 "projects beyond the mirror." But the oblique ray O_1P issuing from O_1 indicates how the observer at B would see the image of the point O_1 appear at the point I_1.

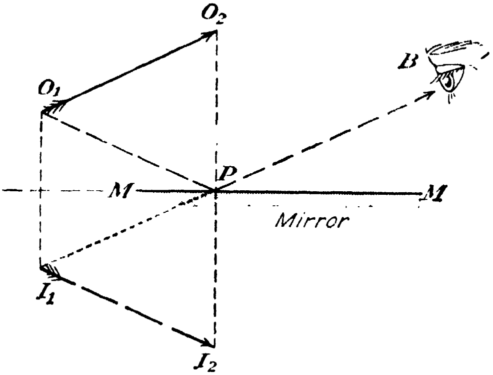


Fig. 30.4.

The images I_1 and I_2 can also be constructed by considering the reflection of two representative rays issuing from each object point. For instance, the two rays issuing from point O_1 in Fig. 30.5 continue to diverge after reflection. Their backward prolongations converge toward a point which is the image point I_1. An observer into whose eye the reflected rays enter will have the impression that they emanate from point I_1. This construction is the same as the one shown in Fig. 30.3, except that, in the latter, one of the two rays issuing from O_1 (the ray O_1I_1) is perpendicular to the mirror.

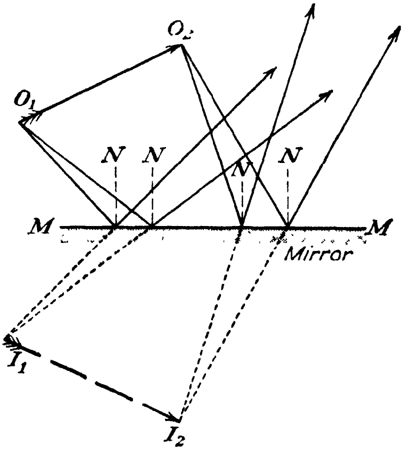


Fig. 30.5. Formation of a virtual image by a plane mirror.

This method of construction is of general importance; it can be used in connection with more complex optical instruments.

The Parabolic Mirror

We have seen that plane waves remain plane (Fig. 26.14) and that spherical waves remain spherical (Fig. 30.2) after reflection from a plane mirror. In certain physical applications, it is important, however, to distort the shape of the wave front, for instance, by converting plane waves into spherical or nearly spherical ones and vice versa. Such a need arises, for example, in connection with the observation of stars. Because of the great distance of stars, the wave fronts of starlight which reach us are practically plane and the density of energy flow is very small. In order to detect a faint star we attempt to "funnel" the energy flux, collecting it from a large area and concentrating it as nearly as possible at a point. Figure 30.6 indicates what we attempt to achieve. The parallel wave fronts W_1 to W_4 move to the right and reach at D a device which bends the wave fronts into spherical shells concave toward the right. The direction of propagation of the wave at any point is perpendicular to the wave front at that point, *i.e.*, in the direction of the indicated radii. As a result it shrinks as shown by the consecutive snapshots of the progressing wave front W_5 to form the fronts W_6, W_7, etc., until it finally shrinks to a point at F which we call the **focus.**

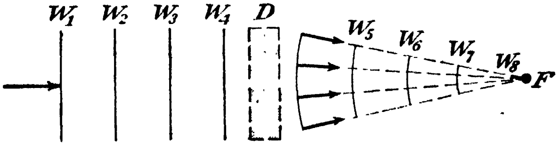


FIG. 30.6. Deformation of plane waves into spherical waves.

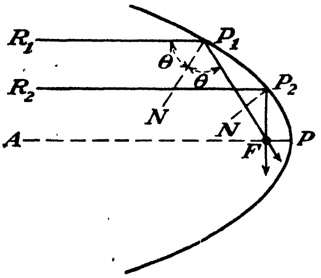


FIG. 30.7. Parabolic mirror.

Such concentration of light energy can be accomplished by reflection of parallel light from a **parabolic mirror.** Its action can be understood more easily by considering a diagram of rays rather than waves. Figure 30.7 shows a cross section of the mirror, the spatial appearance of which can be visualized if we imagine the parabola rotated about the midline AP.

The ability of a parabolic mirror to gather parallel light at one point is due to the following property of the parabola. There is a point F (the focus) on the line AP such that, if we draw *any* line through F, for example, FP_1, and then a parallel P_1R_1 to the axis of symmetry AP through the point of intersection P_1 of this line with the parabola, then the normal N erected at this point bisects the angle between the two lines R_1P_1 and P_1F. This is, however, the familiar construction of the reflected light ray P_1F to the given incident ray R_1P_1, since the action of the parabolic mirror at the point P_1 can be treated as that of a very small plane mirror tangential to the parabola. Similarly, R_2P_2 can be considered as an incident ray of light parallel to the axis AP of the curved mirror which reflects this ray toward the focus F. Thus *all the light entering the wide opening of the reflector is concentrated by reflection at one point F.* The same consideration holds for any kind of waves, so that sound waves, for instance, can also be concentrated ("focused") by the same arrangement.

We realize at the same time that the parabolic mirror can be used to obtain an excellent point source of light, since the original parallel rays all pass through the point F and proceed from there as if they had issued from a point source of light located at F.

Experience shows that *the path of light is reversible,* so that, if we know, for instance, that light coming from R_1 toward P_1 is reflected toward F, we can predict that light coming from F toward P_1 will be reflected toward R_1. Hence, *if a point source of light were put at F, its rays, after hitting the surface of the parabolic mirror, would be converted into parallel rays.*

What practical use can be made of this effect? We know that the intensity of illumination diminishes as the inverse square of the distance from a point source. It would require an extremely intense source to illuminate distant objects. We see, however, that the rays of a point source at the focus of a parabolic mirror do not diverge after reflection from the mirror surface and hence no "dilution" of energy takes place. This effect makes parabolic reflectors invaluable in the design of searchlights, flashlights, automobile headlights, as well as projectors of other types of waves, such as sound and radio waves.

It will be shown below that a similar concentration of light can be accomplished by a spherical mirror (one of circular contour) provided the mirror area is small compared with its radius of curvature. A spherical mirror must have a radius of curvature $R = 2f$ in order to have a focal distance f. The **focal distance** of the parabolic mirror is the distance $PF = f$ of Fig. 30.7.

The Ellipsoidal Mirror

Plane waves can be considered to be small sections of spherical wave fronts of an infinitely distant point source. We can therefore say that a parabolic mirror gathers at one point light which comes from another, infinitely distant point. Is it possible to collect at one point light which comes from a point source at a finite distance? It is possible to do it strictly with a mirror of elliptical contour, owing to the following property of the ellipse. In Fig. 30.8 there are two points on the axis P_AP_B, the so-called foci F_A and F_B, such that *lines drawn from any point of the* *ellipse (e.g., P_1 or P_2) through the two foci form an angle which is bisected by the normal N.* Thus we can imagine F_AP_1 to be a light ray aiming toward P_1 which is reflected toward F_B, and similarly, a light ray F_AP_2 will pass through F_B after reflection. *All rays of light emanating from one focus will pass through the other focus after reflection.* The same consideration can also be applied to sound. A whisper at F_A could be heard at F_B, a large distance from F_A, if a room were designed in ellipsoidal shape (whispering gallery). In optics an ellipsoidal mirror is occasionally used when very faint light coming from a point source is to be concentrated at a distant point with a minimum of loss of radiation.

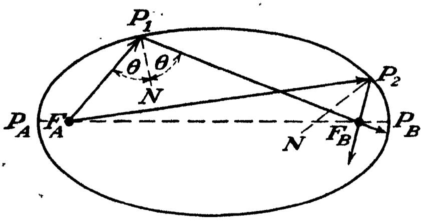


FIG. 30.8. Concentration of light emitted from one focus at the other focus by an ellipsoidal mirror.

Like the parabolic mirror, an elliptical mirror may be replaced by a spherical one if we limit ourselves to a small portion of the mirror in the vicinity of the apex P_B. The focusing is, however, not so sharp in this case as with an ellipsoidal mirror.

The Concave Spherical Mirror

The ellipsoidal mirror has a serious limitation. It can form an image point at one focus only if the point object is located at the other focus of the ellipsoid. As subsequent applications will show, however, it is

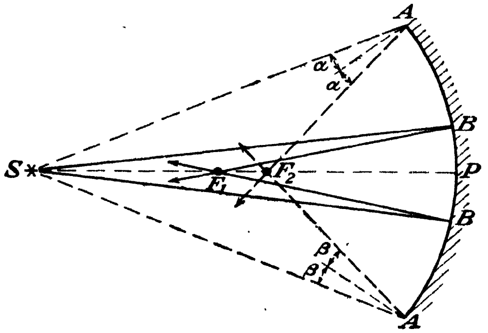


FIG. 30.9. Spherical aberration of a concave spherical mirror.

desirable to be able to focus rays coming from points located at various distances from a curved mirror. This can be accomplished, to a degree of approximation satisfactory in practice, by means of a concave spherical mirror. Figure 30.9 shows a reflection of light, issuing from the point source S, by a concave spherical mirror. For each incident ray the reflected ray can be constructed according to the law of reflection. The normal is simply a radius of the circle in the cross-sectional diagram. Such a construction shows that rays impinging near the edges A of the mirror are united after reflection at the point F_2 whereas rays impinging close to the mirror center P at points between B and P are all focused very nearly at one point F_1. Rays impinging between A and B are focused between F_1 and F_2. Thus the usefulness of such a mirror as a focusing device appears very doubtful. But the focusing ability of the mirror can be greatly improved if we limit ourselves to a small aperture by, say, painting the mirror with nonreflecting paint between A and B or by screening this ring-shaped portion by a diaphragm. As we just stated, *rays reflected from the central portion of the mirror are united approximately at the same common point* F_1. The rays which aim toward F_2 and diverge after passing this point would spoil the point image at F_1 by

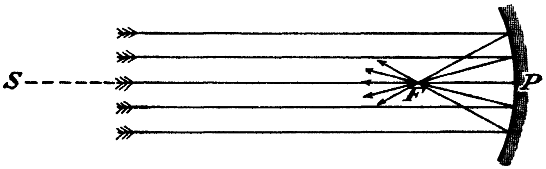


FIG. 30.10. Focusing of parallel rays by a concave spherical mirror.

producing a luminous circle about this point if a screen were held at F_1. This inability of a spherical mirror *of large aperture* to produce sharp point images is called **spherical aberration.** It can be reduced by a diaphragm which limits the aperture of the mirror.

To the observer, the point F_1 will look like the original point source of light, since the light rays seem to emanate from it. It is therefore called an *image of the point source.* The relatively sharp image F_1 of the point source S produced by a mirror of small aperture is formed by rays *actually* intersecting at F_1.

An image formed by actual intersection of light rays is called a **real image** in contradistinction to a **virtual image,** which is obtained by the intersection of the backward prolongation of light rays, such as the image produced by a plane mirror (Fig. 30.5).

In the following considerations of spherical mirrors, we shall limit ourselves to mirrors of small apertures whose spherical aberration is negligible.

The location of the image point F_1 depends on the location of the object S, as can be inferred from a simple geometrical construction of the image points for different positions of S. F_1 moves toward the mirror as S moves away from the mirror toward infinity. For an infinitely distant object S the rays coming from S are all parallel (to the *principal axis SP* of the mirror), and the image is formed at a point F which we shall designate as the focus of the mirror (Fig. 30.10). **The focus is the**

point at which rays parallel to the principal axis are united (just as in the case of the parabolic mirror). As we have mentioned in connection with the replacement of a parabolic mirror by an equivalent spherical mirror, *the focal distance is equal to half the radius of curvature:*

$$\boxed{f = \frac{R}{2}} \tag{30.1}$$

This can be proved as follows (Fig. 30.11): C is the center of curvature of the mirror; hence, the line PC is a normal. Ray AP, which is a parallel to XM (the principal axis), forms the angle of incidence θ with the normal. The angle of reflection CPF is also θ. The angle PFM is equal to 2θ, since the angles APF and MFP are "alternate angles" formed by intersection of two parallel lines with the ray PF, and similarly angle MCP = angle $APC = \theta$.

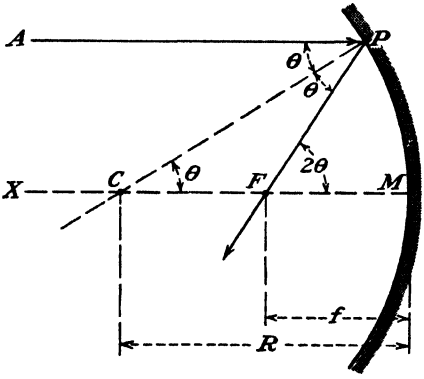


Fig. 30.11.

In Fig. 30.11 the angle 2θ has been made large for clarity. Let us, however, imagine that we are dealing with a mirror of such small "angular aperture" (2θ very small) that the arc PM can be considered practically a straight line perpendicular to XM.

Then from the triangle PCM: $\tan\theta = \dfrac{PM}{CM} = \dfrac{PM}{R} \simeq \theta$ (*a*)

and from the triangle PFM: $\tan 2\theta = \dfrac{PM}{MF} = \dfrac{PM}{f} \simeq 2\theta$ (*b*)

(since because of the smallness of the angles $\tan\theta \simeq \theta$ and $\tan 2\theta \simeq 2\theta$). Thus, from equations (*a*) and (*b*) follows

$$2\theta = \frac{PM}{f} = 2\frac{PM}{R}$$

and hence,
$$f = \frac{\cancel{PM}}{2}\frac{R}{\cancel{PM}} = \frac{R}{2}$$

which was to be proved.

Geometrical Image Construction. We shall now develop a method of constructing geometrically the image of an extended luminous object focused by a concave spherical mirror. Our construction will be based on the fact that light issuing from a point object S is reunited *practically* in one point if S is not too far off the *principal axis* (the line drawn through the center of the mirror M and the center of curvature C).

The arrow O in Fig. 30.12 represents our extended object. Let us construct the images of the tip and the tail of this arrow. We shall assume that light emanates in all directions from every point of the arrow. Since all rays emanating from an object point (for instance, point B) are again united at one image point (b), we have merely to find the intersection of two representative rays in order to locate the image point. Of the multitude of rays issuing from B, we select two whose paths after reflections by the mirror can easily be traced. For instance, we know that the "axis-parallel ray" B-1 is reflected so as to pass through the focus F. Owing to the reversibility of the light path, we know that a "focal ray" F-1 would be reflected as an axis-parallel ray 1-B. Hence, the focal ray BF will be reflected as an axis-parallel ray 2b, and *the image point b is thus located by the intersection of two reflected rays originating at B.*

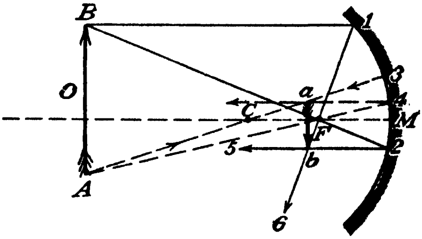


Fig. 30.12. Formation of a real image by a concave spherical mirror.

There are also other rays which can easily be traced; for instance, in locating the image of A we can draw the central ray AC. Since it strikes the mirror normally, it is reflected back along its original path. As our second ray originating from A, we can choose, for instance, the focal ray AF, which is reflected parallel to the axis. The intersection a thus located represents the image point of A, the tail end of the arrow. In similar fashion, an image point can be constructed for every point of the luminous object. We see that *the image is inverted and real* and in this particular case smaller than the object.

If the mirror were to extend merely from point 1 to point 2, the rays B-1 and B-2 would be the limiting rays of the light cone emanating from B and reaching the mirror. The rays b-5 and b-6 would be the limiting rays of the light cone issuing from the image point b. It is clear that an observer could see the image point b only if his eye were disposed within the angle limited by the rays b-5 and b-6. From other positions, this image point would be invisible. If, however, we place a screen of *ground glass* so as to catch the real image ab on it, light is scattered at every image point in *all directions* and the real image can be seen by a large audience without the angular restriction of visibility which has been shown above to exist without the screen.

The Mirror Formula. When we draw the object AB at different distances from the mirror, the preceding construction indicates that the image ab approaches C from the right and increases in size as the object moves toward C from the left. We shall now attempt to derive a mathematical expression relating the image location to the position of the object.

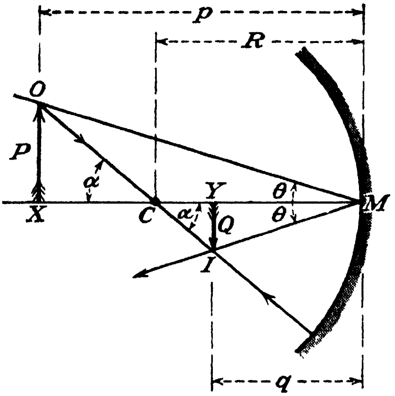


Fig. 30.13.

In Fig. 30.13 only the determination of the image point I of the arrow tip O is shown. The two rays used in this construction are the central ray OC and the apex ray OM. The latter is reflected according to the law of reflection, so that angle OMX = angle XMI. Additional rays serving to establish other image points have been omitted in order not to overload the diagram.

The right triangles CXO and CYI are similar, having the same angles α, 90°, and $(90° - \alpha)$. From the similarity of the triangles follows the proportionality relation

$$\frac{OX}{IY} = \frac{XC}{YC} = \frac{p - R}{R - q} \tag{a}$$

(since $XC = p - R$ and $YC = R - q$).

Furthermore, the right triangles OXM and IYM are similar, having the same angles θ, 90°, and $(90° - \theta)$.

From the similarity of these triangles follows the relation

$$\frac{OX}{IY} = \frac{XM}{YM} = \frac{p}{q} \tag{b}$$

By equating the value given for OX/IY in equation (a) to the value given by equation (b), we obtain

$$\frac{p}{q} = \frac{p - R}{R - q} \tag{c}$$

hence,

$$pR + qR = 2pq \tag{c'}$$

Dividing both sides by pqR we get

$$\boxed{\frac{1}{p} + \frac{1}{q} = \frac{2}{R} = \frac{1}{f}} \tag{30.2}$$

This is the desired **equation for the concave spherical mirror.** We can express its statement particularly simply if we introduce the following terms:

For the reciprocal object distance $\frac{1}{p}$: proximity of object

For the reciprocal image distance $\frac{1}{q}$: proximity of image

For the reciprocal focal length $\frac{1}{f}$: power of the mirror

Equation (30.2) then states: **The sum of the proximities of the image and the object remains constant and is equal to the power of the mirror.** (The object and image distance are measured from the plane drawn at right angles to the principal axis through the point M.)

The magnification M of the image is defined as the ratio

$$M = \frac{\text{image size}}{\text{object size}} = \frac{Q}{P}$$

The following simple relation between magnification, object distance, and image distance can be read off from Fig. 30.13.

$$\boxed{M = \frac{Q}{P} = \frac{q}{p}} \qquad (30.3)$$

which follows from the similarity of the triangles MYI and MXO.

Let us now discuss the physical implications of equations (30.2) and (30.3).

Case *a*. *The object is at infinity.*

Set $p = \infty$ in equation (30.2): $\frac{1}{\infty} + \frac{1}{q} = \frac{1}{f}$

Hence follows: $\underline{q = f}$

Magnification: $M = \frac{q}{p} = \frac{q}{\infty} = 0$

This means that for an infinitely distant object the image shrinks into a point and is located at the focus.

Case *b*. *The object moves from infinity to the center of curvature C.*

When the object is at C $\quad p = R = 2f$

hence $\frac{1}{2f} + \frac{1}{q} = \frac{1}{f} \quad \therefore \frac{1}{q} = \frac{2}{2f} - \frac{1}{2f} = \frac{1}{2f} \quad \therefore \underline{q = 2f = p}$

and $M = \frac{q}{p} = \frac{p}{p} = 1$

In this special case, we obtain a *real, inverted, natural-size image of the object at the same distance from the mirror at which the object is located* (Fig. 30.14). Thus, as the object moves from ∞ to C, the image moves from F (where it is for $p = \infty$) to C and grows in size until object and image meet at C and are equally large.

The real image is inverted and diminished ($M < 1$) for $p > 2f$.

Case *c*. *The object moves from C to F.*

When the object is at F, $\qquad p = f$

and we get: $\qquad \cancel{\frac{1}{f}} + \frac{1}{q} = \cancel{\frac{1}{f}}$

hence $\qquad \frac{1}{q} = 0 \quad \text{or} \quad q = \infty$

and $\qquad M = \frac{q}{p} = \frac{\infty}{p} = \infty$

Thus, as the object moves from C toward F, the image recedes from C to infinity and grows to infinite size. *The real image is inverted and enlarged* ($M > 1$) *for* $2f > p > f$.

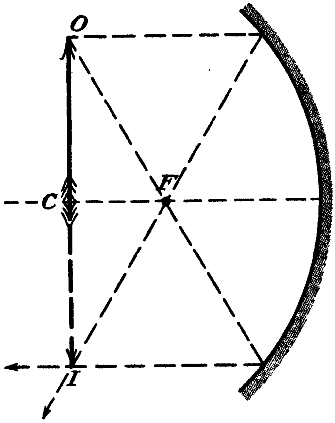


Fig. 30.14. Formation of a natural-size image by a concave spherical mirror.

Case *d*. *The object moves from F toward the mirror.*

It is difficult to visualize what happens to the image when the object moves beyond F. The image reaches infinity and becomes infinitely large when the object reaches the focus. The image cannot move any farther away from the mirror, nor can it grow any larger. We can infer what happens to the image when the object passes the focus by considering equation (30.2): $1/q = (1/f) - (1/p)$. When the object is between F and the mirror, $p < f$ and hence $1/p > 1/f$; hence $1/q < 0$ (*i.e.*, negative). What could a negative image distance mean? We were taking the positive sign of q to mean that q is on the same side of the mirror as the object in the preceding equations. We interpret tentatively the negative sign of q to mean that the image is now behind the mirror.

This interpretation is confirmed by a geometrical construction (Fig. 30.15). By drawing the "central ray" (O-1), the "axis-parallel ray" (O-2), and the reflected rays 1-C and 2-F, we locate the image point I

of the arrow tip as the intersection of the backward prolongation of the reflected rays 1-C and 2-F. The image point lies indeed behind the mirror.

Similarly we can show that, in this case, all image points are obtained by backward prolongation of reflected rays. *The image is consequently a virtual one*, and as can be seen from the diagram, it is *magnified.*

The fact that it is magnified for any object distance smaller than f ($f > p > 0$) can be seen as follows (Fig. 30.15): We draw the ray reflected from the apex M (dotted line) and its backward prolongation, which passes through the image point I. We immediately recognize three equal angles which are marked as θ. From the similarity of triangles OXM and IYM we obtain the following proportion:

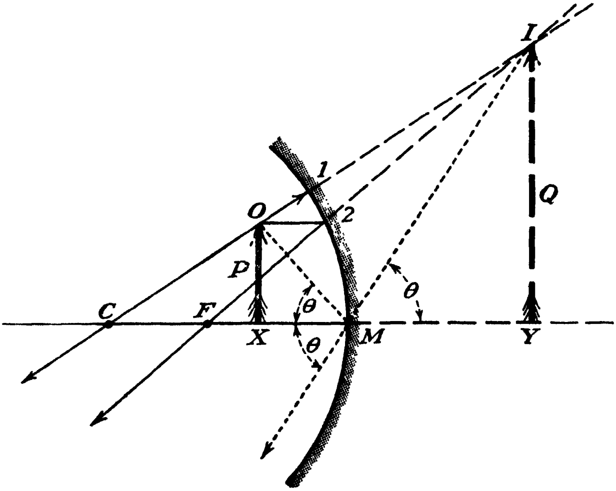


FIG. 30.15. Formation of a virtual image by a concave spherical mirror.

$$\frac{Q}{P} = \frac{YM}{XM} = \frac{q}{p} = M$$

where M is the magnification of the image.

From
$$\frac{1}{q} = \frac{1}{f} - \frac{1}{p} = \frac{p-f}{pf}$$

follows
$$q = \frac{pf}{p-f}$$

Substituting this value of q in the above magnification formula, we get

$$M = \frac{q}{p} = \frac{\cancel{p}f}{\cancel{p}(p-f)} = \frac{f}{(p-f)}$$

We see that, when $p = f$, $M = \infty$.

When p is slightly larger than f, M is very large and positive (since q in $M = q/p$ is positive). When p is slightly smaller than f, M is very large and negative (since q in $M = q/p$ is negative). This means that *the image goes toward* $+\infty$ *as the object approaches* F *from the left and to* $-\infty$ *as the object approaches* F *from the right.* Thus we can say that, as the object passes the point F, the image *jumps* from $+\infty$ (in front of the mirror) to $-\infty$ (behind the mirror), and an infinitely large, real, inverted image is replaced by an infinitely large, virtual, upright image.

It is instructive to visualize these relations by drawing diagrams of a few special cases. Inspection of Fig. 30.15 shows that, as the object approaches the mirror, the paths of the axis-parallel ray and of the corresponding reflected ray do not change but the central ray CO rotates clockwise. As a result its intersection with the backward prolongation of the reflected ray 2-F shifts downward and closer to the mirror. Hence, the virtual image becomes smaller as it moves toward the object.

The magnifying properties of the concave mirror are utilized, for instance, in the shaving mirror and the dentist's mirror. The object is placed in these applications between the mirror and its focus, and its virtual, enlarged, upright image is observed.

We summarize:

The image is real and inverted

a. When $p > 2f$: $0 < M < 1$. The image moves toward the object as the object moves toward the mirror.

b. For $p = 2f = q$: $M = 1$. The image and object meet at this point.

c. For $2f > p > f$: $M > 1$. The image moves away from the mirror as the object approaches the mirror.

The image is virtual and upright

d. For $p < f$: $|M| > 1$. The virtual image moves toward the object and diminishes in size as the object moves toward the mirror.

Convention on Signs

All quantities lying on the nonreflecting side of the mirror are taken as negative, for instance, the distance of the virtual image of any type of mirror and the distance of the virtual focus of a convex mirror (see page 555).

The Convex Spherical Mirror

If the nonreflecting side of the mirror of Fig. 30.13 is polished and thus rendered reflecting, we obtain a so-called *convex spherical mirror.* Such a mirror does not unite axis-parallel rays at one point but rather renders them divergent after reflection. Figure 30.16 serves to show that *axis-parallel rays are reflected so that the reflected rays appear to* *come from a common source point F behind the mirror,* **the virtual focus of the convex mirror,** *whose distance from the apex M is* $f = R/2$.

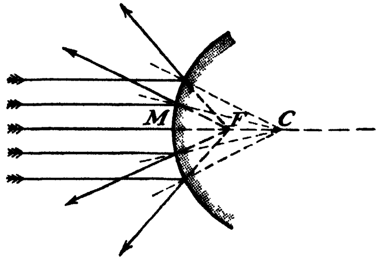


Fig. 30.16. Illustration of the virtual focus of a convex spherical mirror.

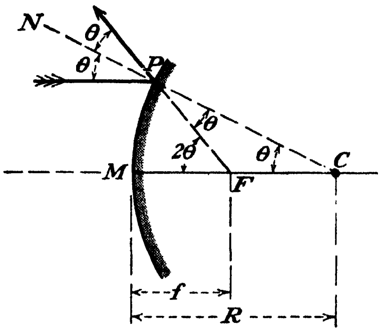


Fig. 30.17.

The relation $f = R/2$ follows from Fig. 30.17. The derivation and diagram are similar to the consideration relating to Fig. 30.11. *For a mirror of very small aperture* we get the following trigonometric relations:

$$\tan\theta = \frac{MP}{R} \simeq \theta \qquad \text{and} \qquad \tan 2\theta = \frac{MP}{f} \simeq 2\theta$$

hence,

$$2\theta = \frac{MP}{f} = 2\frac{MP}{R} \qquad \text{and} \qquad f = \frac{R}{2}$$

Since F lies on the nonreflecting side of the mirror, we consider f a negative quantity according to our convention:

$$\boxed{f = -\frac{R}{2}} \qquad (30.4)$$

Figure 30.18 illustrates the construction of an image formed by a convex mirror. We draw the "central ray" OC, which is reflected back in itself, since it impinges normally upon the mirror. Then we draw the "axis-parallel ray" OA, which is reflected so as to appear to come from the focus F. The intersection of the backward prolongations of these two rays locates the virtual image point I *We obtain an upright, virtual, diminished image.*

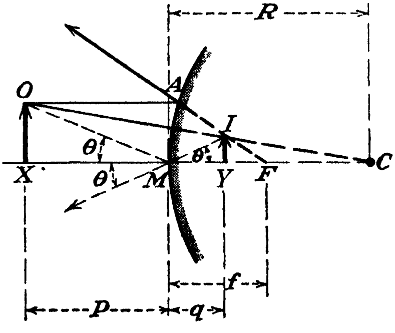


Fig. 30.18. Formation of an image by a convex spherical mirror.

The derivation of the image formula for the convex mirror is practically identical with the derivation of equation (30.2) for the concave mirror. From the similar triangles CXO and CYI of Fig. 30.18 follows

$$\frac{OX}{IY} = \frac{XC}{YC} = \frac{p + R}{R - q} \tag{a}$$

Now we draw the light ray OM and the reflected ray. The backward prolongation of the latter must pass through I, the image point of O, as do all rays which issue from O and are reflected by the mirror. From the similar triangles OXM and IYM follows

$$\frac{OX}{IY} = \frac{XM}{YM} = \frac{p}{q} \tag{b}$$

By equating the value given for OX/IY in equation (a) to the value given by equation (b) we obtain

$$\frac{p}{q} = \frac{p + R}{R - q} \tag{c}$$

hence,

$$pR - qR = 2pq \tag{c'}$$

dividing both sides by pqR, we get

$$\frac{1}{|p|} - \frac{1}{|q|} = -\frac{2}{|R|} = -\frac{1}{|f|}$$

This is the desired relation between image and object distance of the convex spherical mirror. The brackets about the symbols indicate that p, q, R, f stand for the magnitude of the quantity regardless of its sign. If we adopt the convention of taking the distance of all virtual images with the negative sign, we can write the above equation as follows:

$$\boxed{\frac{1}{p} + \frac{1}{q} = \frac{1}{f}}$$

This is identical with the equation for the concave mirror. *When applying this expression practically to a convex mirror, we substitute the numerical values for q and f with a negative sign.* f is also taken negative, since the focus can be considered as the virtual image of an infinitely distant point object.

It is easy to see from Fig. 30.18 that *the image is virtual and diminished for all positions of the object* and remains on the nonreflecting side of the mirror. As we move the object XO toward the mirror, we do not change the path of the "axis-parallel ray" OA and of the reflected ray upon whose backward prolongation lies the point I. Point I is located at the intersection of this backward prolongation with the line OC. As object OX moves toward M, this line becomes steeper and the point of intersection I moves higher and closer to the mirror. Thus the image remains always behind the mirror and moves toward the object as it approaches the mirror. At the same time the size of the image approaches the size of the object.

The magnification of the image follows from the similar triangles XOM and YIM:

$$M = \frac{IY}{OX} = \frac{q}{p}$$

It is the same expression as for the convex mirror, but in this case $M < 1$ for all values of p.

Lenses

The ability of spherical mirrors to distort the shape of an incoming wave front enables them to create real or virtual images of extended objects. In all cases of image formation by spherical mirrors, the observer has to be on the same side of the mirror as the object. This is a great handicap in many practical applications. It is desirable to develop a means of accomplishing the same thing with the aid of a device which could be held between the observer and the observed object. Such a device is a **spherical lens.** It can be deduced from the law of refraction that a transparent body bounded by smooth spherical surfaces (a spherical lens) immersed in a medium of a refractive index different from its own distorts the shape of a wave front in a similar fashion to a spherical mirror. A plane wave, for instance, can be converted into a spherical wave approximately converging toward a point (Fig. 30.19*a*) or diverging from a virtual focus (Figs. 30.19*b* and *c*). In Figs. 30.19*a* and *c*, the refractive index of the lens material is higher than that of the surroundings; in Fig. 30.19*b* it is lower.

We distinguish between **converging and diverging lenses.** The action of a converging lens is similar to that of a concave mirror. It focuses parallel light at one point (Fig. 30.19*a*). The action of a diverging lens is analogous to the action of a convex mirror, converting plane waves into spherical waves appearing to originate in a virtual focus (Fig. 30.19*c*). We shall limit the following discussion to a consideration of lenses in a vacuum. Figure 30.20*a* illustrates the shapes of lenses which act as converging lenses in a vacuum. The lenses shown in Fig. 30.20*b* act as diverging lenses in a vacuum.

What general criterion enables us to distinguish between a converging and a diverging lens?

A converging lens is thickest at the center.

A diverging lens is thinnest at the center.

It is easy to understand how the converging power of a lens is influenced by its shape. In Fig. 30.19*a* the plane wave front entering the lens from

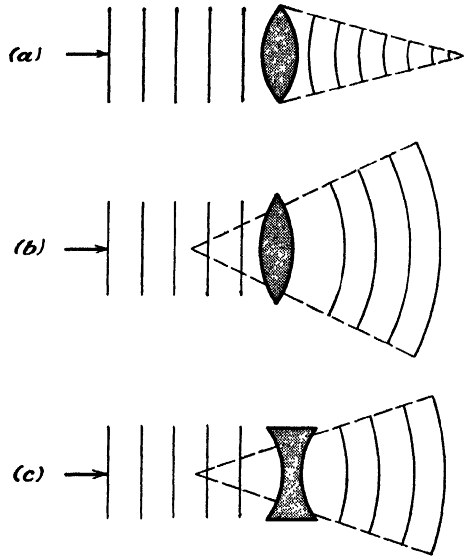


FIG. 30.19. Deformation of plane waves by lenses. *a.* Biconvex lens in an optically less dense medium. *b.* Biconvex lens in an optically denser medium. *c.* Biconcave lens in an optically less dense medium.

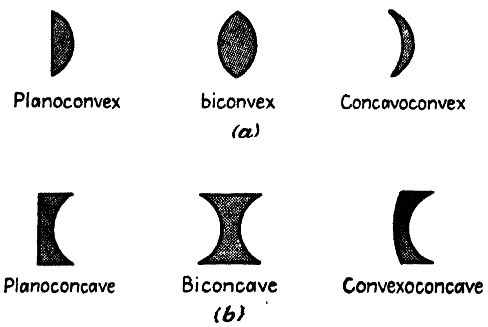


FIG. 30.20*a*. Types of converging lenses. *b*. Types of diverging lenses.

the left at the center has to travel through the lens medium (in which the velocity of light is smaller than in the surrounding medium) a longer distance than in other portions of the lens. Hence, this portion of the wave front will lag behind those portions which passed through the lens closer to the rim, and the wave front will be distorted as shown in Fig. 30.19*a*.

In Fig. 30.19*c*, on the other hand, the center portion of the wave front will traverse the shortest distance in glass and will be retarded less than the portions of the wave passing through the peripheral regions of the lens. This explains the different type of distortion of the wave front transmitted through a biconcave lens.

In Fig. 30.19*b* light is assumed to travel faster through the lens than through the surrounding medium; hence, the wave portion traveling through the lens center is retarded least and comes out ahead of the peripheral portions. Hence, the lens acts as a diverging lens.

The Lens Equation. The theory of lenses is considerably more complicated than the mirror formulas and will be omitted here. We shall give without proof the *lens equation* which can be derived for *thin lenses* *i.e.*, lenses whose thickness is negligible as compared with the radii of curvature of their surfaces. The form of the lens equation is identical with the mirror equation, and its physical meaning will be discussed below. The **lens equation** is

$$\boxed{\frac{1}{p} + \frac{1}{q} = \frac{1}{f}} \qquad (30.5)$$

In words: **The refractive power of a lens** $(1/f)$ **is equal to the sum of the proximities** $[(1/q)$ and $(1/p)]$ **of image and object,**†

where
$$\frac{1}{f} = (n - n')\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \qquad (30.6)$$

p and q are the object and image distance from the lens, respectively; f is the focal distance, *i.e.*, the distance from the lens center to the point at which a *converging lens* unites parallel light or from which originally parallel light waves appear to diverge after passing through a *diverging lens*. n is the refractive index of the lens material, and n' that of the surrounding medium ($n' = 1$ for vacuum and nearly so for air). r_1 and r_2 are the radii of curvature of the two refracting lens surfaces. By taking the radius of a convex surface of the refractive medium to be positive and that of a concave surface of the refractive medium to be negative, we achieve that the value of $[(1/r_1) + (1/r_2)]$ is negative for lenses which are thinner at the center and positive for those which are thicker at the center. Thus *the focal length of a con-*

---

† The refractive power is measured by opticians in **diopters.** It is defined as the reciprocal focal length expressed in meters; *e.g.*, a lens of $f = 2$ m is a lens of ½-diopter refractive power.

verging lens is positive and of a diverging lens negative. This assignment of signs must be observed when equation (30.5) is used for both converging and diverging lenses.

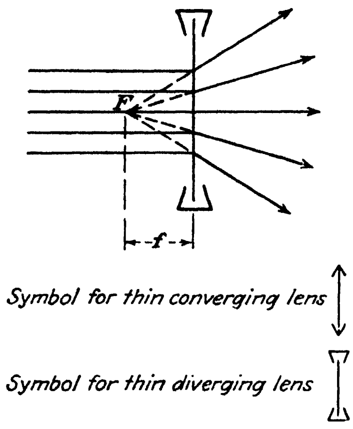

F
f
Symbol for thin converging lens
Symbol for thin diverging lens

FIG. 30.21. Illustration of the virtual focus of a diverging lens.

We adopt the following general convention on signs in the lens equation:

All virtual distances (*i.e.*, distances from the lens of points defined by the intersection of the backward prolongation of rays) **are taken as negative.**

This includes, as we shall see, the following distances measured from the lens center:

1. Distance of the virtual focus F of a diverging lens (Fig. 30.21).
2. Distance of the virtual image produced by a converging lens.
3. Distance of the virtual image produced by a diverging lens.

The Converging Lens. We shall now discuss the action of a converging lens. We shall develop a geometrical image construction based on the following idealized properties of a *thin* converging lens:

1. Rays parallel to the principal axis AB, which is defined as the line passing through both radii of curvature of the lens surfaces, are refracted so as to pass through one point, the focus F (Fig. 30.22*a*). Hence, conversely, rays passing through the focus are refracted so as to proceed parallel to the principal axis. (Reverse the direction of rays in Fig. 30.22*a*.)

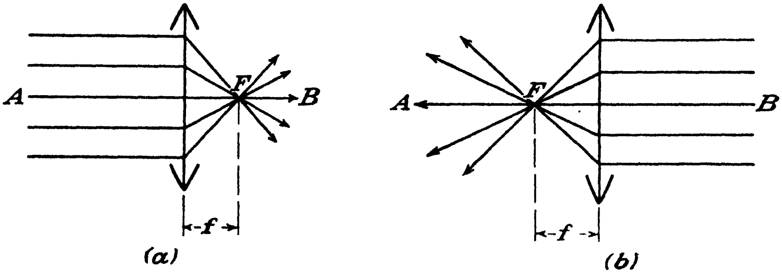


FIG. 30.22*a*. Focusing of parallel light rays by a converging lens. *b*. The focal distance is not changed when the direction from which the light comes is reversed.

2. The focal distance is the same for light coming from either side of the lens (Figs. 30.22*a* and *b*).
3. A ray passing through the lens center is not deviated.

The reason for property 3 can be seen as follows: A ray passing through a thick slab of glass is refracted as shown in Fig. 30.23. It is displaced laterally but does not change direction. The thinner the plate the smaller is the lateral displacement. The center portions of the two surfaces of a thin lens can be considered to be parallel and very close to each other; hence, the deviation of the ray passing through the lens center can be neglected.

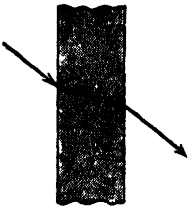


Fig. 30.23. Refraction of light by a plane parallel glass slab.

4. Light rays issuing from an object point which is located off the principal axis are united in a distinct image point.

Actually our basic assumption 1 is only roughly correct. We get **spherical aberration** in the case of a lens as well as in the case of a spherical mirror. The rays passing near the periphery of the lens are united closer to the lens than the rays passing closer to the center (Fig. 30.24). This leads to a diminution of the sharpness of an image. The spherical aberration can be reduced as in the case of a mirror by a diaphragm which screens the rim of the lens. (It is also possible to improve the image by using the rim rays and cutting out the central ones.)

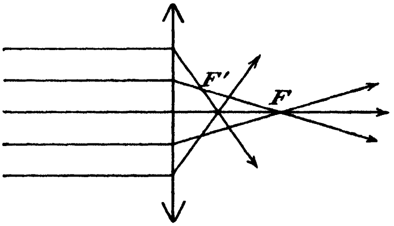


Fig. 30.24. Spherical aberration of a converging lens (greatly exaggerated).

In addition to spherical aberration, lenses possess another defect when they are used with white light, namely, **chromatic aberration.**

Figure 30.25 shows rays W of mixed light impinging upon a lens. The color mixture is red and violet. Owing to the fact that the index of refraction of the lens is higher for violet light than it is for red light (dispersion), the violet rays are refracted more strongly and are united in a focus F_V which is closer to the lens than the focus F_R of the less refrangible red rays. As a result, if we adjust a lens to give a sharp violet image, the red light will not be united in a sharp point in the

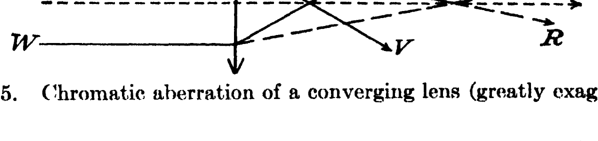


Fig. 30.25. Chromatic aberration of a converging lens (greatly exaggerated).

plane of the violet image and a blurred image will be obtained with mixed or white light.

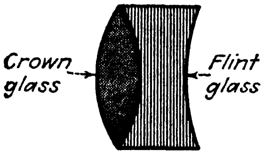


FIG. 30.26. An achromatic lens combination.

Chromatic aberration can be eliminated by the use of monochromatic light, or it can be annulled for two extreme wave lengths of light by an appropriate combination of lenses, *e.g.*, a converging lens with a diverging lens. The latter must be made of a material whose refractive indices for the two colors differ from those of the material of the converging lens. (The combination of crown glass and flint glass is commonly used for this purpose.) Such **achromatic lens combinations** are basic elements of optical instruments. Figure 30.26 shows such an "achromat." The focal length f of the lens combination for a given color is obtained from the focal lengths f_1 and f_2 of the individual lenses according to the equation

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \quad \text{(for closely spaced thin lenses)}$$

(where f_2, the focal distance of the diverging lens, is negative). The values of f_1 and f_2 are chosen so that the focal length of the combination is the same for red as it is for violet light. The action of the diverging lens can be described as follows: The converging lens unites mixed parallel light in two foci as shown in Fig. 30.25. The diverging lens moves the violet focus more effectively away from the converging lens than it moves the red focus. The lenses are chosen so that the violet focus catches up with the red focus so that both coincide.

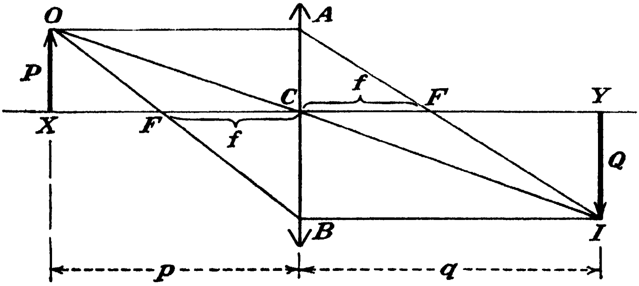


FIG. 30.27. Formation of a real image by a converging lens.

In the following discussion lenses are assumed to be free of spherical and chromatic aberration. Figure 30.27 illustrates the graphical determination of the image point I for the object point O. Again, as in the case of the mirror, we can select from the cone of rays which emanate from point O and pass through the lens two representative rays whose intersection localizes the image. In Fig. 30.27 three rays have been drawn to indicate the possibilities in the selection of representative rays. The central ray OC passes without deviation. The focal ray OB becomes parallel to the axis after refraction, and the "axis-parallel ray" OA becomes a focal ray. Any pair of these rays would have been sufficient to locate the image point I. The image construction can be carried out point by point. In the present case the image YI turns out to be real, inverted, and somewhat enlarged. The magnification is seen to be $M = Q/P = q/p$, since from the similar triangles OXC and IYC it follows that $Q/P = q/p$.

Let us review now how the image changes as the object moves from infinity toward the converging lens:

Case *a*. *The object is at infinity.*

Set $p = \infty$ in equation (30.5): $\frac{1}{\infty} + \frac{1}{q} = \frac{1}{f}$

Hence follows $\underline{q = f}$

The magnification is $M = \frac{q}{p} = \frac{q}{\infty} = 0$

This means that the image of an infinitely distant object shrinks into a point and is located at the focus.

Case *b*. *The object moves from infinity to the distance $p = 2f$.*

When $p = 2f$,

$$\frac{1}{2f} + \frac{1}{q} = \frac{1}{f} \qquad \therefore \frac{1}{q} = \frac{1}{f} - \frac{1}{2f} \qquad \therefore \frac{1}{q} = \frac{2}{2f} - \frac{1}{2f} = \frac{1}{2f}.$$

$$\underline{q = 2f = p} \qquad \text{and} \qquad M = \frac{q}{p} = \frac{q}{q} = 1$$

We obtain a *natural-size, inverted, real* image when both object and image are at the distance of $p = q = 2f$ from the lens.

Thus as the object moves from infinity to the right toward the lens, the image moves from the right focus to the right, away from the lens. But the object moves much faster. While the image covers the distance of f centimeters in moving from $q = f$ to $q = 2f$, the object covers the distance extending from infinity to $2f$ in chasing the image.

When $\infty > p > 2f$, $M = \frac{q}{p} < 1$, since $q < p$. *The image is diminished for $p > 2f$.*

Case *c*. *The object moves from $p = 2f$ to F.*

For $p = f$ we get $\frac{1}{q} = \frac{1}{f} - \frac{1}{p} = \frac{1}{f} - \frac{1}{f} = 0$

hence, $q = \infty$ and $M = \frac{q}{p} = \frac{\infty}{p} = \infty$

As the object moves from $p = 2f$ toward F, the image increases to infinite size and moves to infinity. Now the image moves faster than the object, covering the distance from $2f$ to infinity while the object covers the distance f.

For $2f > p > f$: $$M = \frac{q}{p} > 1 \text{ since } q > p$$

As the construction shows, *the image remains real and inverted in all cases mentioned above.*

Case *d.* *The object moves into the space between the lens and the focus.*

In this case $p < f$. $$\frac{1}{q} = \frac{1}{f} - \frac{1}{p} = \frac{p - f}{pf} \qquad \therefore q = \frac{fp}{p - f}$$

Magnification is $$M = \frac{q}{p} = \frac{pf}{p(p - f)}, \qquad M = \frac{f}{p - f}$$

From $1/q = (p - f)/pf$ we see that q is *negative*, since $p < f$. According to our convention, a negative q indicates a *virtual image.*

Since $|p - f| < f$ as long as $p < f$, it follows that

$$M = \frac{f}{p - f} > 1$$

This means that the virtual image is magnified for all values of $p < f$.

Figure 30.28 shows the construction of the magnified virtual upright image in the case of $p < f$. The image point I is located by the backward prolongation of the ray AF and of the central ray OC. An observer at B would have the impression of seeing an object at the intersection of the prolonged rays. Since the object remains between the image and the lens for $p < f$, q remains $> p$ and hence $M > 1$. From the diagram we see that $Q/P = q/p > 1$.

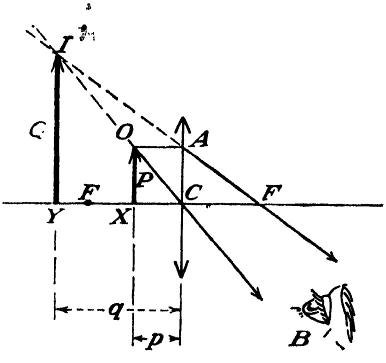


FIG. 30.28. Formation of a virtual image by a converging lens.

(The student should show that, as the object passes the left focus, the image jumps from plus infinity to minus infinity; *i.e.*, the real image at infinite distance on the right of the lens is replaced by a virtual image at infinity on the left of the lens.)

Summary. The image remains real and inverted as the object moves from infinity toward the lens until it reaches the distance $p = f$. The magnification remains <1 until the object reaches the distance $p = 2f$.

From there it increases toward an infinite value which is reached for $p = f$. The image moves in the same direction as the object. As the object passes the focus, the image jumps from $+\infty$ to $-\infty$ and becomes virtual and upright. $|M|$ remains >1 for $p < f$ and diminishes as p decreases.

In all cases the magnification formula $M = q/p$ is valid.

The Diverging Lens. We shall limit ourselves to a brief discussion of the construction of the image formed by a diverging lens and of the application of equation (30.5) to the case of a diverging lens. Figure 30.29 illustrates the formation of the virtual image of the arrow tip O. As we have stated before, rays parallel to the principal axis XX' are refracted so that their backward prolongation passes through a common virtual focus which lies on the object side of the lens. This enables us to draw the broken ray O-1.

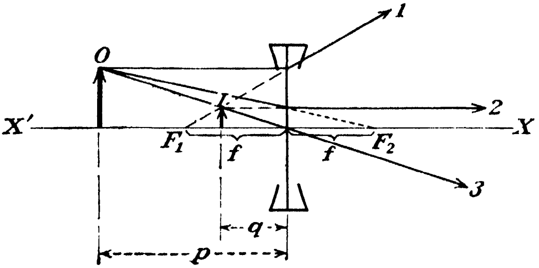


FIG. 30.29. Formation of a virtual image by diverging lens.

If we trace ray 1 in reverse, we see that it aims toward F_1 and is refracted toward O, becoming parallel to the axis. *All rays aiming toward a focus become parallel to the axis after refraction.* This knowledge enables us to draw the broken ray O-2.

Finally, this being a "thin lens," we can draw an undeviated central ray O-3. The backward prolongations of rays 1, 2, and 3 intersect at point I, which is the virtual image of point O. Only two of these rays are necessary to determine the image point.

It is clear that I, the image point of O, will always lie on the central ray, regardless of the distance p of O from the lens. It is also clear from the diagram that the intersection of the prolonged ray 1-F_1 with the central ray O-3 will be always above the axis and between the object and the lens. Hence for all object distances, $q < p$. We see thus that *the image is always upright, diminished, and virtual.*

For $p = \infty$ we get $\frac{1}{q} = \frac{1}{f} - \frac{1}{p} = \frac{1}{f}$, $q = f$, and $M = \frac{q}{\infty} = 0$

This is the mathematical expression of the fact that the virtual focus is the virtual image of an infinitely distant object.

According to our convention on signs in the lens equation, *we always have to use a negative sign for q and f with a diverging lens*, since the image formed by a diverging lens is always virtual. This includes the focus as the image of an infinitely distant point.

OPTICAL INSTRUMENTS

We shall now review how the properties of lenses and mirrors have been utilized in various optical instruments. We shall disregard details, limiting ourselves to the discussion of principles only.

The Photographic Camera

The pinhole camera (Fig. 28.3) can be used for photographic recording of an image by placing a photographic plate in place of the ground glass G.

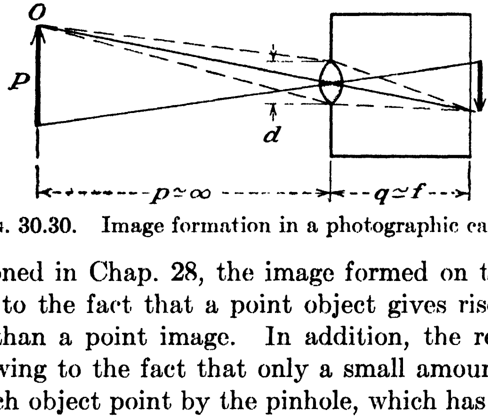


FIG. 30.30. Image formation in a photographic camera.

As was mentioned in Chap. 28, the image formed on the plate is quite imperfect due to the fact that a point object gives rise to an elliptical image rather than a point image. In addition, the required exposure time is high owing to the fact that only a small amount of light is collected from each object point by the pinhole, which has to be very small for reasonable image definition. These difficulties are remedied by replacing the pinhole by a large opening which is filled out by a converging lens (Fig. 30.30). The dashed rays issuing from point O illustrate how light entering the lens gives rise to a point image for each point of the object.

In most uses of a camera, the object distance is very much greater than the focal length of the lens, so that the image is formed close to the focal plane. If we were, for instance, to double f in Fig. 30.30, the image would be formed twice as far from the lens. Its linear dimensions would be doubled, and its area would be quadrupled. If we imagine for simplicity that the object is a square of area P^2, its original image area being equal to Q^2, then it is easy to see that, after doubling the focal distance (and hence the length of the camera), the final image area will be $(2Q)^2 = 4Q^2$. If the total flux of light energy from the object entering the camera is ϕ, the original average illumination of the plate over the image area is ϕ/Q^2. After doubling the focal length, the illumination becomes $\phi/4Q^2$. Thus, we see that the illumination E on the plate is approximately inversely proportional to the square of the focal length under conditions illustrated in Fig. 30.30 (where $f \simeq q$):

$$E \propto \frac{1}{f^2} \tag{a}$$

If we were to double the lens diameter, the area of the base of each cone of light entering the camera would be about quadrupled (the conical bundle of light issuing from O and entering the camera is indicated by dashed lines in Fig. 30.30). Thus the total flux of light which enters the camera and is distributed over the image area is about quadrupled. It is thus easy to see that the average illumination of the plate is approximately proportional to the square of the diameter of the lens:

$$\underline{E \propto d^2} \tag{b}$$

Hence, the relation between the average illumination of the plate, the focal length, and the diameter of the camera lens can be represented by the following expression:

$$E \propto \left(\frac{d}{f}\right)^2 \tag{c}$$

The "speed" of a camera is proportional to the illumination E and is usually indicated by the so-called "f number" f/d, which is inversely proportional to $\sqrt{E}$. An $f/4.5$ lens is one whose diameter is 1/4.5 of its focal length.

By considering the cone of light rays entering the lens from each object point and the change of the image size with object distance, the student should attempt to answer the question whether the required exposure time is reduced by bringing a distant object closer to the camera.

A simple lens is not a satisfactory camera "objective" owing to various defects such as spherical and chromatic aberration, astigmatism (inability to form sharp point images for object points deviating appreciably from the central axis), and other lens deficiencies. Some correction of these imperfections can, however, be accomplished by an appropriate combination of converging and diverging lenses of various kinds of glass. To show the complexity of a typical camera objective, Fig. 30.31 illustrates the well-known Zeiss Tessar lens.

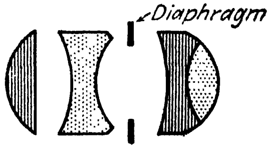


FIG. 30.31. A compound camera objective (Zeiss Tessar).

A camera objective is usually equipped with an *iris diaphragm,* an opening of variable diameter which can be adjusted so as to screen a desired portion of the peripheral section of the lens, thus diminishing the effective lens diameter. It is desirable to use as small a diaphragm opening as possible (still giving reasonable illumination on the plate) in order to obtain maximum sharpness of the image. (Why does narrowing the diaphragm opening increase the sharpness of the image?)

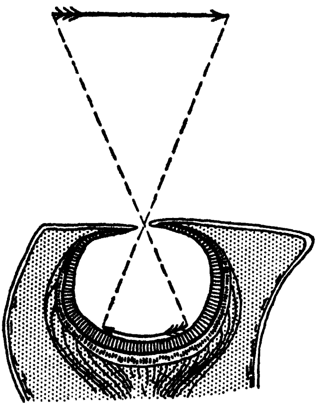


FIG. 30.32. Cross section of the eye of the nautilus.

The adjustment of the camera to different object distances is accomplished by varying the distance between the plate and the lens.

The Eye

One of the most primitive eyes is the lensless eye of the molluscan creature Nautilus. This eye is essentially a pinhole camera (Fig. 30.32).

The human eye is designed similarly to a lens camera (Fig. 30.33). C is a transparent membrane (cornea) covering the front portion of the eye. Light entering the *cornea* passes through media of different refractive index: the chamber A, filled with *aqueous humor* ($n \simeq 1.336$); the (inhomogeneous) *crystalline lens* L ($\bar{n}_{\text{av.}} \simeq 1.437$); and the gelatinous *vitreous humor* V ($n \simeq 1.336$). The image is formed on the light-sensitive layer R (retina) which contains nerve endings (*rods* and *cones*), the rods distinguishing only between light and darkness, the cones discriminating between different colors.

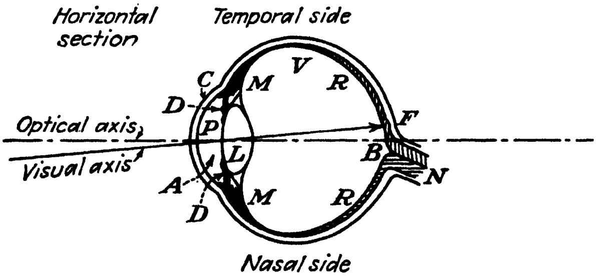


FIG. 30.33. Cross section of the human eye.

Owing to the fact that the indices of refraction of the transparent sections of the eye differ very little, the main refraction takes place at the cornea. The image is formed on the retina. The function of the crystalline lens is merely that of fine adjustment in the process of focusing the image on the retina. The lens can be deformed by the ring-shaped *ciliary muscle M*. When the latter contracts, the lens assumes a more spherical shape, thus increasing the refractive power of the optical system. When the lens is in a state of relaxation, the image of an infinitely distant object is formed on the retina of a normal eye. When the eye comes close to the object, the image shifts backward beyond the retina. The contraction of the ciliary muscle, however, deforms the lens so that the image is shifted forward onto the retina. This process is called **accommodation.**

The *iris diaphragm D* in front of the crystalline lens plays the same role as the diaphragm of a camera objective. By contracting, it limits the effective diameter of the eye lens and thus the amount of light entering the eye. It responds automatically to an increase of illumination by contraction.

The lens system of the eye is far from being perfect. It exhibits spherical and chromatic aberration. In doing work requiring visual distinction of fine details, it is desirable to minimize spherical aberration. This can be achieved by illuminating the object very intensely. The intense light entering the eye stimulates the *pupil P* (opening of the iris) to contract and thus diminishes spherical aberration.

At the point *F* of the retina there is a little groove (*fovea centralis*) which is marked by a yellow spot surrounding it. This area is distinguished by a high concentration of cones. *Vision is sharpest when the image of the point of attention is focused on the fovea.* In viewing details the eye automatically turns so as to catch the desired image on the fovea centralis.

The entrance *B* of the optic nerve interrupts the retinal network of nerve endings. This spot is insensitive to light and is called the *blind spot.* (Can you devise a simple experiment to demonstrate the existence of the blind spot?)

Some Common Defects of Vision. For the normal eye a range exists in which the object must be located in order to allow the eye to form a sharp image of it on the retina. The limits of this range are called the *far point* and the *near point.* The far point of a normal eye is at infinity.

In an abnormally long eyeball the image of an infinitely distant object is formed in front of the retina. Accommodation of the eye lens cannot shift this image back to the retina, but on the contrary, contraction of the ciliary muscle moves it even closer to the pupil. The object can be seen sharply by such an eye only if it is brought close enough to it.

Such an eye is said to be *nearsighted* (*or myopic*). The near point of a nearsighted eye is closer to the eye than the near point of a normal eye.

If the eyeball is abnormally short, the image of an infinitely distant object is formed behind the retina (*farsighted, or hyperopic, eye*). The eye has to accommodate in order to see at a great distance. The accommodation required for a near object is greater than for a normal eye, since the image must be produced closer to the pupil in the hyperopic eye. As a result the near point is farther from the eye than in the normal case.

Another frequently encountered defect of vision is *astigmatism*. This abnormality is due to variation in curvature of the cornea. For instance, if we make a horizontal and a vertical section of an astigmatic cornea, we shall find two different curvatures of this main refractive surface of the eye. An astigmatic eye when looking at millimeter paper can accommodate to see sharply either the vertical lines or the horizontal lines but not both at the same time.

Eye Glasses. The inequality of curvature of the **astigmatic eye** can be corrected by placing a cylindrical lens in front of the eye. If, for instance, the vertical section through the cornea shows too large a curvature, a diverging cylindrical lens whose vertical section is curved (say planoconcave) but whose horizontal section is straight is placed in front of the eye; in the case of too small a curvature, refraction is aided by a converging cylindrical lens.

The problem of a **farsighted eye** is the following: The normal "distance of distinct vision" (*i.e.*, the average distance at which people with normal eyes would hold a book) is about 25 cm. For a farsighted eye the near point may be, say, 1 m, which is inappropriate for reading. What could be done to enable a farsighted man whose near point is 100 cm from his eye to read a book held at the normal distance of 25 cm? We can solve this problem by making use of the ability of a converging lens to form an upright virtual image *behind the object* (see Fig. 30.28). The eye on the right side of this lens sees the virtual image I. We can choose the image distance q equal to the near point of the hyperopic eye to enable it to see the near object P held at a conveniently close distance. In our case, $p = 25$ cm (distance at which the book is to be read). $q = -100$ cm (desired distance of the virtual image). What should be the focal length f of the prescribed converging lens?

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f} \qquad \therefore \frac{1}{25} - \frac{1}{100} = \frac{1}{f}$$

$$\frac{1}{f} = \frac{4}{100} - \frac{1}{100} = \frac{3}{100} \qquad \therefore \underline{f = 33.3 \text{ cm}}$$

The solution of the problem of the **myopic eye** is achieved in a similar fashion. Suppose that the far point of a nearsighted person is 200 cm from his eye. What could we do to enable him to see stars?

We could make use of the property of a diverging lens to create an upright virtual image of a distant object between the object and the eye (Fig. 30.29). The lens can then be chosen so that the virtual image of the distant object is no farther from the myopic eye than its far point, and then the image of the object will be seen clearly.

In our case the answer is very simple. The object distance is $p = \infty$. The distance of the image should be at most $q = -200$ cm (q is negative because the image is virtual).

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f} \qquad \therefore \frac{1}{\infty} - \frac{1}{200} = \frac{1}{f}$$

$$\underline{f = -200 \text{ cm}}$$

Thus, a diverging lens of $\frac{1}{2}$ diopter should be prescribed for this eye.

The Magnifier. The myopic eye possesses one advantage over the normal eye. Owing to the closeness of the near point to the eye a nearsighted person can bring an object much closer to his eye than a normal person. He thus obtains a larger retinal image of the object, which enables him to distinguish finer detail. A person with normal vision can accomplish the same by using an auxiliary lens between his eye and the object, the latter being located between the eye and its near point. The lens creates in this case a virtual magnified image *beyond the near point*, enabling the observer to obtain a sharp retinal image. This case is illustrated in Fig. 30.28. By changing the distance between the object P and the lens of a given focal length, we can, for instance, make the image Q appear at the "distance of distinct vision," *i.e.*, at $q = -25$ cm from the lens which we assume to be held close to the eye.

In this case, $$\frac{1}{p} = \frac{1}{f} - \frac{1}{q} = \frac{1}{f} - \frac{1}{-25} = \frac{1}{f} + \frac{1}{25}$$

Hence the magnification is

$$\underline{M = \frac{|q|}{|p|}} = 25\left(\frac{1}{f} + \frac{1}{25}\right) = \underline{\frac{25}{f} + 1} \qquad (30.7)$$

The magnification is seen to increase with decreasing focal length.

The magnifying lens is generally used so that the eye is placed near point C right behind the lens. Figure 30.28 shows that the angles subtended at C by the image and by the object are the same.

The Compound Microscope

Considerably higher magnifications than with a simple magnifier can be achieved with the *compound microscope.* Its principle is based on the use of an **objective** lens L_1 (Fig. 30.34) to create a real magnified image of the tiny object O and of an **eyepiece** lens L_2 which serves as a magnifying glass, *i.e.*, creates an enlarged virtual image I_2 of the real image I_1 formed by the objective L_1.

The total magnification M is equal to the product of the magnification m_1 of the objective and the magnification m_2 of the eyepiece. The distance between the object O and the objective is very nearly equal to f_1, since in practice $L \gg f_1$ (*i.e.*, the image X_1-I_1 can be practically considered as an infinitely distant one).

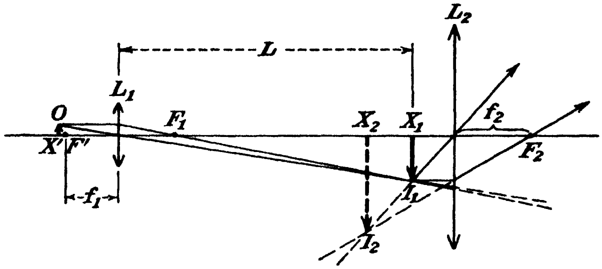


FIG. 30.34. Image construction for a compound microscope.

Hence approximately $$m_1 \simeq \frac{L}{f_1}$$

and $$m_2 = \frac{25}{f_2} + 1$$

[if the virtual image X_2-I_2 is formed 25 cm from L_2, see eq. (30.7)]. Hence the magnification of the compound microscope is†

$$\boxed{M = m_1 m_2 = \frac{L}{f_1}\left(\frac{25}{f_2} + 1\right)}$$

L is practically equal to the length of the microscope tube.

At high magnifications the light forming an image is spread over a large area; hence, very intense illumination of the object is required to give a sufficiently bright image. This is accomplished by a special *condensing lens* system which converges a powerful light flux upon the object. In addition, in order to avoid reflective losses of light, cedar oil is used with some types of objectives as an optical medium between the microscope slide cover and the objective. The index of refraction of this oil is practically equal to that of glass. Thus, there is no reflection at the glass-oil interface (**oil-immersion** system).

---

† In practice one forms the visual image X_2-I_2 at infinity. In this case, $m_2 = 25/f_2$.

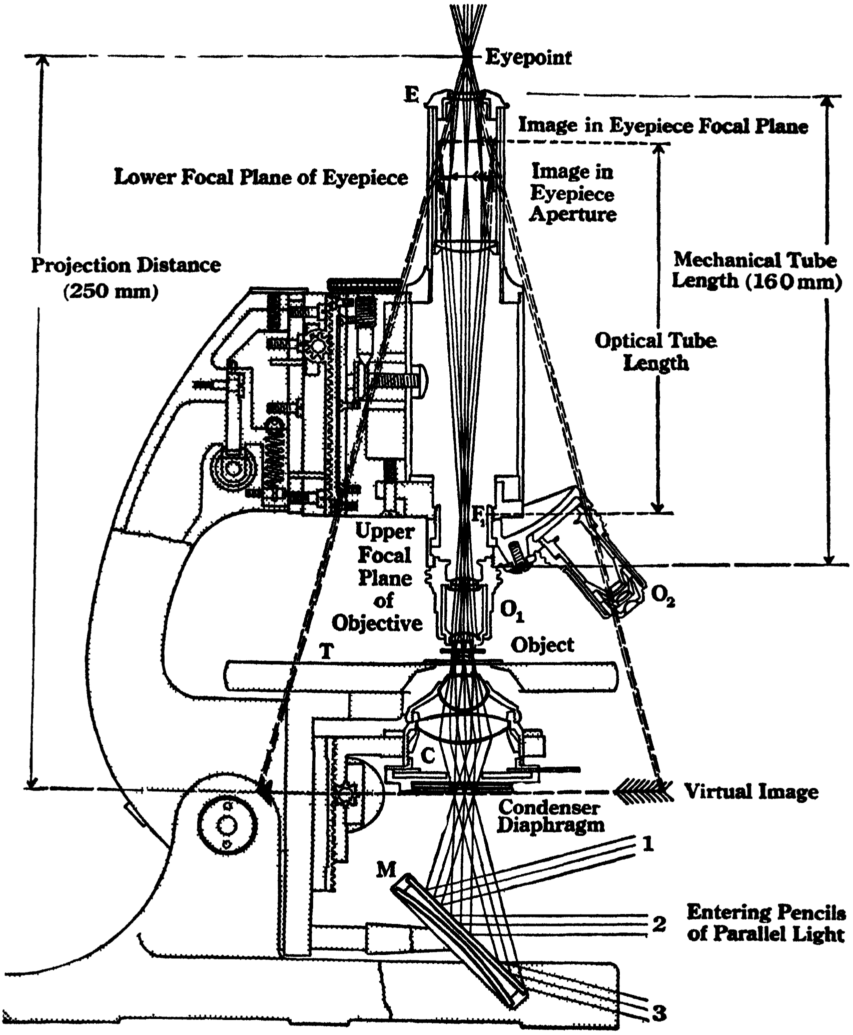
In modern microscopes the objective is a complicated system of lenses corrected for various aberrations, and the eyepiece is itself a (relatively simple) compound microscope. The focal length of the objective is very much smaller than that of the eyepiece.

Figure 30.35 shows a sectional view of a modern microscope. M is a mirror which reflects light into the condenser C. T is a table provided with a small opening through which the light coming from the condenser can pass. The object under investigation is placed over this opening. O_1 is the objective. O_2 is an objective of a different power which can be rotated into the position of O_1 to replace it. E is the eyepiece.

In recent years a revolutionary invention has been made in the field of microscopy: **the reflecting microscope.** It is no exaggeration to say that it is the most promising development in microscopy since the invention of the compound microscope. Interestingly enough, this "revolutionary innovation" was known to Newton. He built reflecting telescopes and also designed a reflecting microscope, using mirrors in place of lenses. But this line of development was not pursued in microscope making after one had succeeded in preparing satisfactory achromatic objectives and eyepieces. The reflecting microscope owes its revival to the extension of the range of radiations which can be used for microscopy. As we shall see in Chap. 31, there is an advantage in using the invisible ultraviolet and infrared radiation for observation. But an objective which is achromatic for visible light will, in general, not be achromatic for these radiations. In addition, ultraviolet radiation is very strongly absorbed by glass. Quartz lenses can be used instead of glass optics, but this solution leaves much to be desired.

The reflecting microscope (Fig. 30.36) consists essentially of a concave mirror, which produces an enlarged real image of the object, and is achromatic for all types of radiation, since the law of reflection applies to radiations of all wave lengths. It can be focused with visible light and the illumination changed to an invisible radiation, such as ultraviolet light, and the image produced on the photographic plate will remain sharp. We shall point out in Chap. 31 why this achromatic feature makes this microscope a revolutionary development. Here we shall limit ourselves to a consideration of its design.

Figure 30.36 shows the optical system designed by Burch. The combination of the concave mirror Mc and the convex mirror mc acts as a *condenser*, concentrating the light which enters from below through the opening in Mc upon the object at point O which rests on the horizontal


Eyepoint
E
Image in Eyepiece Focal Plane
Lower Focal Plane of Eyepiece
Image in Eyepiece Aperture
Mechanical Tube Length (160 mm)
Projection Distance (250 mm)
Optical Tube Length
F_1
Upper Focal Plane of Objective
O_2
O_1
T
Object
C
Virtual Image
Condenser Diaphragm
1
M
2
Entering Pencils of Parallel Light
3

Path of Light through the Microscope with standard Abbe Condenser

FIG. 30.35. Cross-sectional view and light path of a modern microscope. (*Courtesy of Bausch and Lomb Company.*)

transparent table. The combination of the convex mirror *mo* and concave mirror *Mo* acts as an *achromatic objective.* Light emanating from any point of the object *O* is reflected from the mirror *Mo* toward the mirror *mo,* which reflects the rays toward the opening in the mirror *Mo.* We can imagine that the emerging rays converge slightly and unite at a different point on a photographic plate, which is not shown in the diagram. Conical bundles of light emanating from different points of the object *O* are united at different points of the photographic plate on which the enlarged real image is projected.

Among the advantages of the reflecting microscope is its greater "working distance" (up to 20 mm) between the object and objective as compared with the working distance of a conventional microscope of comparable optical characteristics (less than 1 mm). This advantage is important in micromanipulation and in microscopic observations of tissues in vivo. Thus organs such as brain, liver, and kidney left in their normal place can be examined on living, anesthetized animals under high magnification.

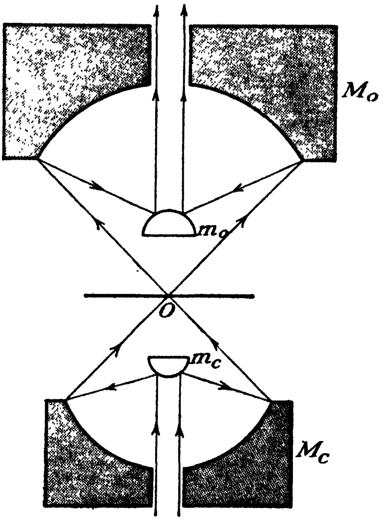


FIG. 30.36. Schematic diagram of a reflecting microscope. (*From Barer, in Lancet, Mar.* 26, 1949, *p.* 533.)

The reflecting microscope can do all that a refracting microscope can do and many other things in addition which the latter cannot accomplish.

When we look at biological material, such as a living cell, through a microscope, we notice that it is optically inhomogeneous. There are particles which appear dark, consisting of material which absorbs light more strongly than the surrounding fluid, and in many types of cells there are colored granules consisting evidently of material which absorbs light of some wave lengths more than of others. There may also be visible enclosures, such as oil globules, of an index of refraction differing markedly from that of the surroundings. As such enclosures, however, become smaller, and as their index of refraction approaches that of the surrounding fluid, their visibility diminishes. A great gain in contrast can be obtained under such conditions by the use of **dark-field illumination.** The principle of this method is illustrated by the commonly remarked fact that dust particles, which are not noticeable in a brightly lit room, become clearly discernible in a dark room into which a thin beam of sunlight enters through a small hole in the window shade. The dust particles entering the beam of light become sources of diffracted and reflected light and can be seen as luminous points against the dark background when one looks at the beam of light from the side.

Figure 30.37 shows a *dark-field condenser* which illuminates a microscopic object so that the illuminating light rays do not enter the microscope objective. Objects, such as Spirochaete pallida, which cannot be seen with the conventional transillumination can be made visible in a similar fashion to the dust particles of the above example, through dark-field illumination. A microscope equipped with a dark-field condenser is referred to as an **ultramicroscope.** It owes this name to the fact that it renders visible colloidal particles which are too small for observation in a conventional microscope. The resolving power of ordinary microscopes is not great enough to show the shape of such particles. They appear as tiny disks surrounded by diffraction rings moving about at random, performing a lively Brownian motion.

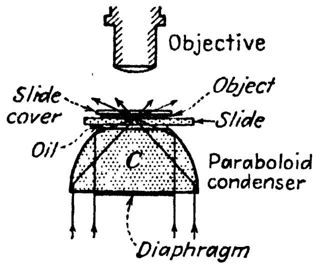


FIG. 30.37. Dark-field illumination by means of a paraboloid condenser (C). (In use the objective is much closer to the object and the space between objective and slide cover is filled with oil.)

Microscopy has been enriched in recent years by another method of observation, **phase microscopy,** which is of very great interest to biologists. This method makes it possible to see transparent enclosures in a medium of only slightly different index of refraction which cannot be made visible by conventional methods of observation. Thus, for instance, the motion of the chromosomes in mitotic or meiotic cell division can be observed in a living cell by means of a phase microscope.

Phase microscopy makes use of the fact that light waves passing through a transparent particle of a refractive index slightly above that of the surroundings suffer a slight retardation in phase. The transmitted wave of retarded phase can be conceived of as being composed of the original incident wave plus a superimposed diffracted wave retarded by $\lambda^*/4$ with respect to the incident wave. Through an ingenious arrangement which cannot be discussed here the diffracted and nondiffracted waves are separated from each other and the phase of the latter is retarded further so as to increase the phase difference between the two waves to $\lambda/2$. When the two waves are reunited, they are in phase opposition and the transparent particle appears dark on a light background. By making the phase difference λ instead of $\lambda/2$, the par-

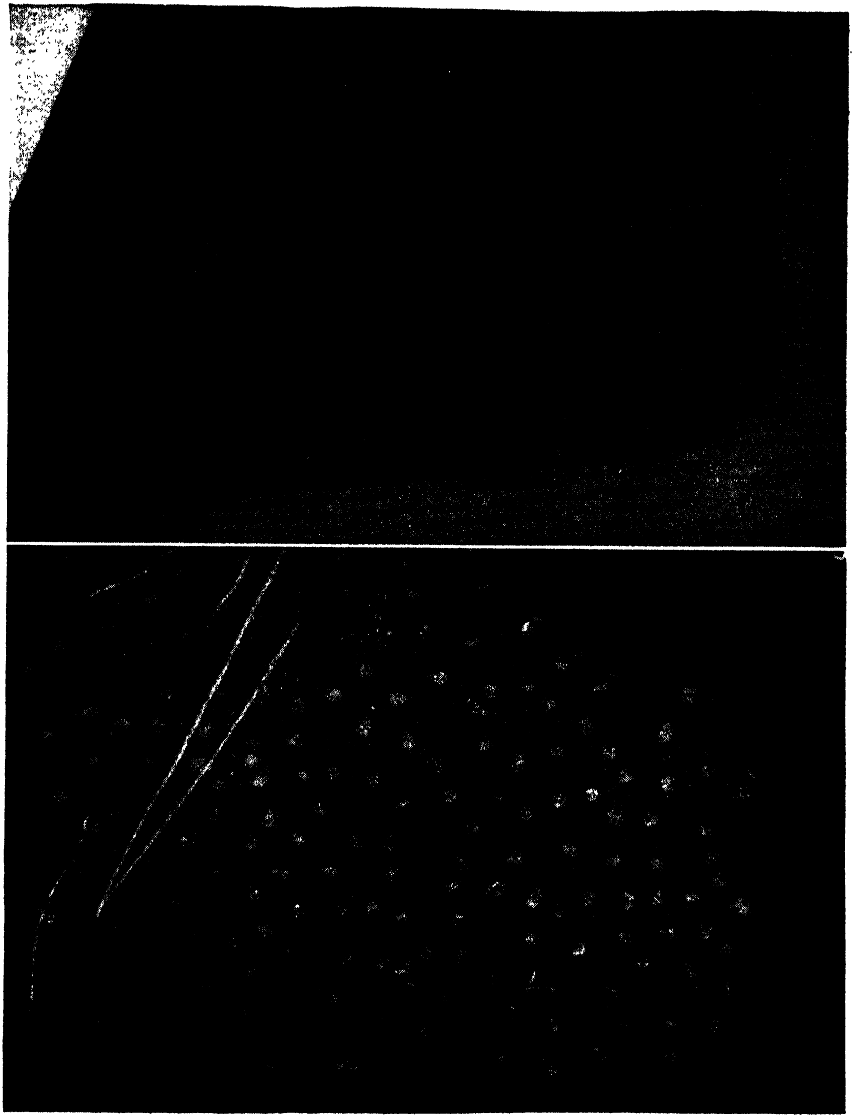


FIG. 30.38. Microphotograph of the epithelium of frog's nictitating membrane. (*Above*) Photograph taken with a conventional microscope. (*Below*) Photograph taken with a phase microscope. (*From Bennett, Jupnik, Osterberg, Richards, in Trans. Am. Microscop. Soc., p.* 119, 1946.)

ticles can be made to appear brighter than the background. Figure 30.38 illustrates the effectiveness of phase contrast microscopy. Figure 30.38*a* is a microphotograph of the epithelium of frog's nictitating membrane taken with an ordinary microscope. Figure 30.38*b* shows the same tissue photographed through a phase microscope.

Telescopes

The purpose of the microscope is to create a large virtual image of a minute object which forms too small a retinal image when viewed by the unaided eye. Large objects will also give too small a retinal image when viewed from a large distance. The optical instrument (telescope) designed to increase the angle which a distant object subtends at an observer's eye (and thus to increase the retinal image) is similar in design to a compound microscope. Figure 30.39 shows a *Keplerian telescope.* It consists, like a microscope, of an **objective lens** L_1 and an **eyepiece** lens L_2.

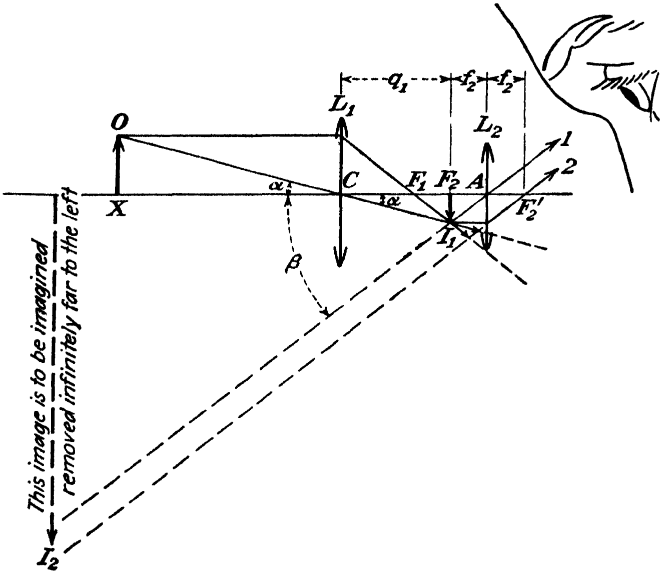

q_1
f_2
f_2
L_1
L_2
O
X
C
α
α
F_1
F_2
A
1
2
F_2'
I_1
β
I_2
This image is to be imagined removed infinitely far to the left

FIG. 30.39. Image construction for a Keplerian telescope.

The objective lens L_1 forms a real image I_1 nearly in its focal plane† if the object is very far away. Thus, in practice, $q_1 \simeq f_1$ ($f_1 = CF_1$, see diagram). This real image is viewed with a magnifying glass (objective

---

† In this diagram the focus F_1 and the image I_1 have been drawn far apart for clarity. Actually the size of the image is to be imagined much smaller and its location is to be pictured very close to F_1. The foci F_1 and F_2 practically coincide.

L_2) in such a way that the virtual image I_2 appears at infinity. This means that the real image I_1, which forms the object of the eyepiece L_2, must be in the focal plane of L_2. Under these conditions the ray issuing from I_1 parallel to the axis is refracted so as to run parallel to the central ray (see rays 1 and 2). The backward prolongations of these rays "meet at infinity." This locates the virtual image point I_2 at infinity. The infinitely distant image I_2 is indicated in the diagram where it should not be. I_2 should be drawn at the intersection of the dashed rays very (theoretically infinitely) far away.

The advantage of using a telescope is apparent from Fig. 30.39. If we were to place our unaided eye at C to view the distant object O, it would subtend the angle α. On the other hand, when we look at it through the telescope, the angle subtended is $\beta > \alpha$.

The angular magnification $M = \beta/\alpha$ is equal to the magnification of the retinal image thus achieved. M can be determined from the diagram as follows:

from the triangle CF_2I_1: $\tan \alpha = \dfrac{F_2I_1}{q_1}$

and since $q_1 \simeq f_1$ (for a distant object), remembering that for small angles $\tan \alpha \cong \alpha$, we get

$$\alpha \simeq \frac{F_2I_1}{f_1} \tag{a}$$

from the triangle AF_2I_1: $\tan \beta = \dfrac{F_2I_1}{f_2} \simeq \beta$ $\qquad$ (b)

Hence, $$\underline{M = \frac{\beta}{\alpha} = \frac{F_2I_1}{f_2}\,\frac{f_1}{F_2I_1} = \frac{f_1}{f_2}} \tag{30.9}$$

The angular magnification of the telescope is equal to the ratio

$$\frac{\text{Focal length of the objective}}{\text{Focal length of the eyepiece}}$$

This accounts for the fact that telescopes are usually so long: The focal length of the objective is chosen very long to achieve a large magnification. The length of the telescope tube is approximately equal to the sum of the focal lengths of the objective and of the eyepiece.

The telescope picture is inverted. It can be reversed by using an intermediate "inverting lens" for purposes of terrestrial observation, or the reversal can be accomplished by special reflecting mirrors or prisms.

One can also use a concave mirror as a telescope objective. This type of an instrument is called a **reflecting telescope,** or a "reflector." Reflectors are widely used by astronomers because of several important advantages over *refractors* (telescopes using an objective *lens*): (1) A parabolic mirror, which focuses parallel star light more perfectly than a spherical lens or a spherical mirror, can be used as an objective. (2) There is no chromatic aberration. (3) Photographically effective ultraviolet rays, which are absorbed by glass, are reflected by the mirror and thus utilized. (4) Only one surface need be ground in making a mirror. (5) A large lens is deformed slowly owing to the pull of gravity and its images become imperfect, whereas the mirror can be supported by a heavy foundation and distortion can thus be avoided. This method of mounting allows us to use mirrors many times as large as the largest feasible lenses. This enables us to photograph fainter stars.

Resolving Power of Optical Instruments

The reason for building astronomical objectives of huge diameter is not alone due to the desire to collect as much light as possible from a faint star. A large objective has a large **resolving power**; *i.e.*, it allows us to distinguish two neighboring stars as two individual objects in cases where a smaller objective may make them appear fused into one.

Figure 30.40 shows a series of photographs taken of four point objects with a diaphragm in front of the objective lens. In Fig. 30.40*a*, the diaphragm is very small and the diffraction patterns are strongly pronounced. The two images on the extreme right are not "resolved." In Fig. 30.40*b*, the diaphragm is larger and the images are just resolved; in Fig. 30.40*c*, the diaphragm is wide open and the images are clearly resolved.

The advantage of making the objective in optical instruments large as compared with the wave length of light can be understood from the following discussion:

1. We have seen in Chap. 29 (see Fig. 29.9) that parallel light passing through a slit is diffracted so as to illuminate an area on a distant screen which is wider than the slit [solve formula (29.5) for x].

2. Similarly, we have seen that a circular hole gives a diffraction pattern analogous to that of a narrow slit (Fig. 29.11). In this case, however, the diffraction pattern consists of a central bright circle surrounded by a series of concentric dark rings alternating with bright ones. The expression for the diameter $2R$ of the first dark ring is similar to the one for the separation of the first dark fringes in the case of a slit:

$$2R = 2.44L \frac{\lambda}{D}$$

where L = distance of the hole from screen, D = diameter of the hole, $2R$ = diameter of the first dark ring (which determines the diameter of

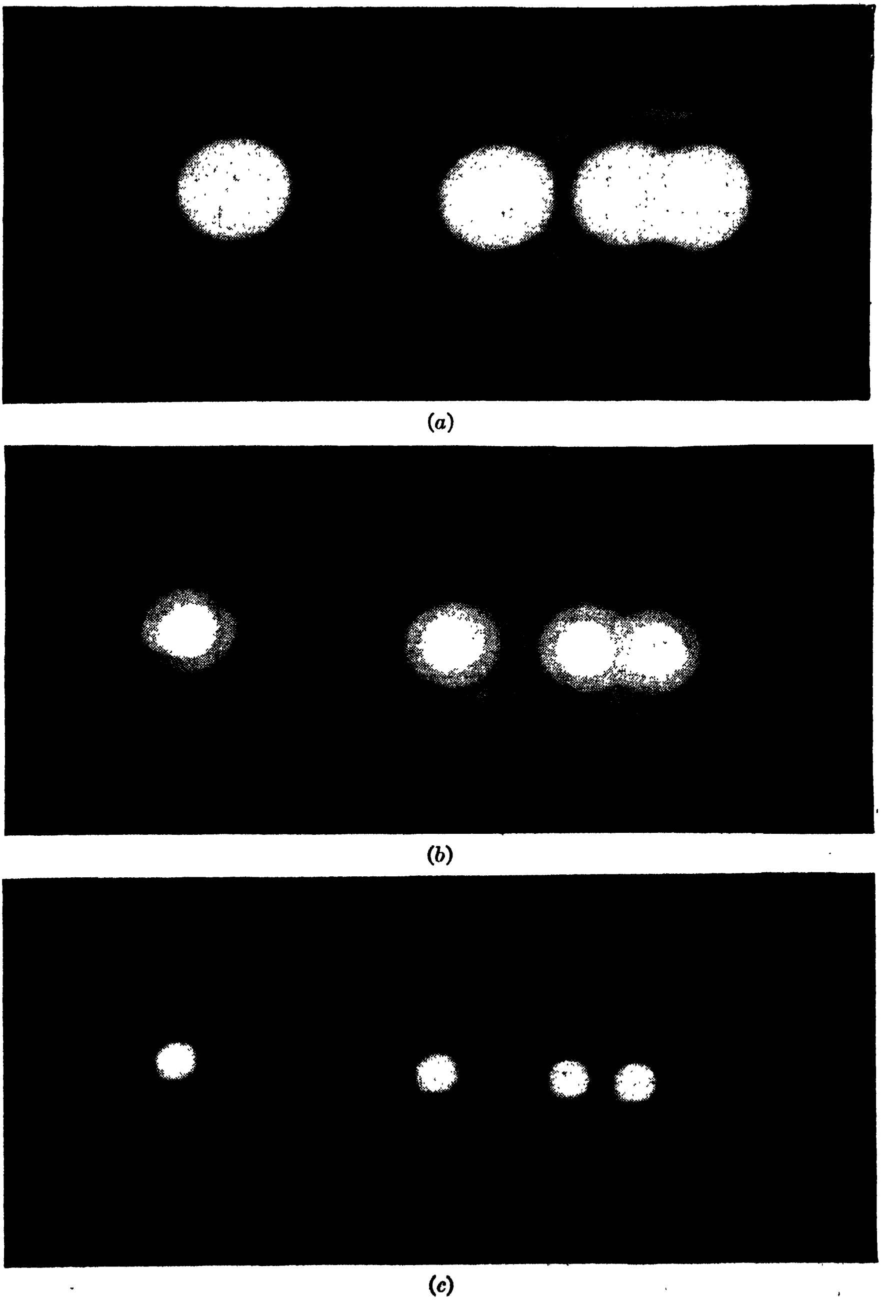


FIG. 30.40. Effect of the diameter of the objective on the resolving power of a camera. (*From Sears and Zemansky, "College Physics," Addison Wesley.*)

the inner bright circle). [Compare the above expression with the following equation which follows from (29.5): $2x = 2L(\lambda/D)$.]

We see from this formula that *the central bright spot decreases in diameter (a) when λ decreases, (b) when D increases.*

3. If in the above-mentioned case the circular aperture were filled by a converging lens, one would expect (according to geometrical optics) the lens to unite parallel rays at one point. Actually, however, owing to the wave nature of light, diffraction takes place also in the case where the aperture is filled by a lens, and a diffraction pattern similar to the one described in paragraph 2 appears on the screen instead of a sharp focal point. In other words, instead of an image point we get an image circle surrounded by alternating bright and dark rings. The diameter $2R$ of the inner bright circle (taken to be equal to the diameter of the first dark ring) depends on the diameter D of the lens and on the wave length of light:

$$2R = 2.44L\frac{\lambda}{D}$$

4. Thus the image formed by the lens is not sharp because of the overlapping of the image circles which we obtain instead of image points. The image could be made sharper by reducing the radius R of an image circle. The above equation shows that this can be achieved in a variety of ways: (*a*) by making the diameter D of the lens large (this is one of the reasons why telescope objectives are made very large in diameter); (*b*) by observing with light of a small wave length (this is utilized in microscopes of high magnification by using blue, violet, or ultraviolet light for illumination); (*c*) it can also be shown that by immersing the object in a medium with a high index of refraction the resolving power of the microscope can be increased (oil-immersion system).

The ability of an objective to produce two separate images of two neighboring object points is a measure of its *resolving power.* The smaller the separation between two object points which can still be distinguished as separate objects the greater is the resolving power. Two point objects are considered as just resolved when the distance between the centers of their diffraction circles is equal to the radius of the first dark ring.

Figure 30.40 demonstrates how the wave nature of light limits the resolving power of optical instruments and thus sets an ultimate limit to our ability to investigate structural detail by means of visible light.

Example

The focal length of the objective of a microscope is $f_1 = 5$ mm and of the eyepiece $f_2 = 50$ mm. Determine the magnification achieved when the object is placed 0.1 mm beyond the focus of the objective. (Treat this objective and eyepiece as thin lenses.)

Solution. The object distance is $p = 5.1$ mm. The image distance q follows from

$$\frac{1}{5.1} + \frac{1}{q} = \frac{1}{5} \qquad \therefore \frac{1}{q} = \frac{1}{5} - \frac{1}{5.1} = \frac{0.1}{25.5}$$
$$\therefore q = 255 \text{ mm}$$

The magnification of the real image produced by the objective is

$$m_1 = q/p = 255/5.1 = 50$$

The magnification of the eyepiece which serves as a magnifying glass to produce a virtual enlarged image at the distance of distinct vision of $q = 250$ mm is

$$m_2 = \left(\frac{250}{f_2} + 1\right) = \left(\frac{250}{50} + 1\right) = 6$$

Hence total magnification is $M = m_1 m_2 = 50 \times 6 = 300$.

QUESTIONS AND PROBLEMS

1. On sunny days perfectly circular patches of light can often be seen in the shade of trees. They must be due to sunlight passing between the leaves. How could you explain the perfect circular shape of these patches of light?

2. Describe the imperfections of image to be expected if one should use a simple converging lens as a camera objective.

3. What is the advantage of a lens camera over a pinhole camera?

4. Does the speed of a camera depend on the focal length of the lens?

5. If an object moves closer to the camera lens, does its image, which is produced on the film, get brighter or less bright? (The brightness of the image should be judged by the average illumination of the film, *i.e.*, light flux per unit area.)

6. What is nearsightedness due to?

7. What is farsightedness due to?

8. What is astigmatism due to?

9. How can the visual defects be corrected?

10. What nerve endings are responsible for color vision?

11. It is generally stated that good illumination is of advantage in doing work which requires visual attention to small details of pattern. Is there a physical reason for that?

12. Microscope objectives have usually a very short focal distance. Why?

13. What is the advantage of an oil-immersion objective?

14. Why are the objectives and eyepieces of microscopes composed of many lenses?

15. What are the advantages of a reflecting telescope ("reflector") as compared with a "refractor"?

16. How could you design a lens which would unite light of all colors as nearly as possible at one focal point (achromatic lens)? What would be the advantage of such a lens?

17. When we look at our mirror image as we raise our right hand, we see it lift its left hand. Right and left seem to be interchanged. Why then don't we see ourselves upside down in a mirror?

18. Can you design a periscope, an instrument to enable the soldiers in a trench to see the battlefield?

19. Can you design a terrestrial telescope in which inversion of the image is accomplished by (*a*) lenses? (*b*) prisms?

20. Can you design an instrument to look into the interior of the eye (ophthalmoscope)?

21. How could you use a microscope to look at stars?

22. How tall, at least, must a mirror be in order to enable a man to see his full figure?

23. A concave mirror of focal length $f = 5$ cm is used as a simple projection microscope. How far from the mirror must the object be placed in order to obtain a hundredfold magnification?

24. A concave mirror produces a real image of a child who is 1 m tall, 5 m away from the mirror, at a distance of 5.05 cm from the mirror. How tall is the image of the child?

25. How far from the face should one hold a concave spherical shaving mirror of focal length $f = 50$ cm in order to obtain a fivefold magnification?

26. A ball bearing of 1 cm diameter is used as a convex mirror. What is its focal length? Determine graphically the image formed by the central portion of this ball bearing of a nail placed 10 cm away from it. Find the magnification.

27. A slide is to be projected by means of a lens of 10-cm focal length upon a wall which is 5 m away from the slide. How far from the slide must the lens be placed? How large is the image if the slide is 5 cm square?

28. A certain camera produces a natural-size image of an object when the distance between the lens and the plate is 20 cm. How large is the focal length of the camera objective? How far is the object from the lens?

29. Find by a *graphical method* the focal length which a lens must have in order to create an image one-fifth of the natural size 60 cm away from the object?

30. How tall is the image of a person 1.5 m tall produced by a camera objective of 5-cm focal length if the person is 2 m away from the lens?

31. The near point of a farsighted person is 50 cm for the left eye and 25 cm for the right eye. Prescribe a pair of glasses to correct for his visual defect.

32. The far point of a nearsighted person is 10 cm for the left eye and 20 cm for the right eye. Describe a pair of glasses to correct for his visual defect.

33. The effective refractive power $(1/f)$ of a combination of two *closely* placed thin lenses is given by the sum of the refractive powers of the two lenses, the refractive power of a diverging lens being taken negative: $1/f = (1/f_1) + (1/f_2)$. You are given a combination of a converging lens of the focal length of 15 cm and a diverging lens of a focal length of 10 cm. Find the distance and size of the image formed by this lens combination of a candle 10 cm tall placed 50 cm in front of the double lens. Is the image upright or inverted?

34. A magnifying glass of 5-cm focal length is held so as to produce a 10-fold enlarged virtual image of an insect. How far is the insect from the lens? How far is the image from the lens? (Assume the eye to be held next to the lens.)

35. An objective of a certain microscope has a focal length of 0.4 mm, and its eyepiece a focal length of 4 cm. The object distance from the *focus* of the objective is 0.2 mm. How far from the objective must the eyepiece be placed if it is used to project a sharp image of the object upon a wall 2 m away from the object? How large is the magnification achieved? (Treat objective and eyepiece as thin lenses.)

36. The focal length of the objective of a telescope is $f_1 = 5$ m. Its angular magnification is $m = 20$. How long is the telescope?

37. Design a reflecting telescope using a parabolic mirror as an objective and a lens as an eyepiece.

38. Design a simple reflecting microscope using a mirror as an objective and a lens as an eyepiece.

BIBLIOGRAPHY

Rayton, W. B., Microscopes, in Glasser's "Medical Physics," 1st ed., pp. 733–750, Year Book Publishers, Inc., 1944.

Barer, R., The Reflecting Microscope, *Lancet*, 1949, p. 533.

Rayton, W. B., Microscopic Dark Field Illumination, in Glasser's "Medical Physics," 1st ed., pp. 744–745, Year Book Publishers, Inc., 1944.

Spencer Optical Co., "Phase Microscopy." (There is a simplified theoretical discussion in this pamphlet on phase microscopes.)

CHAPTER 31

THE STUDY OF SPECTRA

a. What is the acoustical counterpart to a spectroscope?

b. How is (1) a grating spectroscope, (2) a prism spectroscope designed?

c. What are spectral lines?

d. What are Fraunhofer lines?

e. Can you design a spectrograph which will produce spectral circles instead of spectral lines?

f. How could you determine gases present in a gas mixture by means of a spectroscope?

g. How could you determine what elements are present in a powder spectroscopically?

h. How could you perform spectral analysis of a colored solution?

i. In determining elements present in the stars do we make use of the emission or of the absorption spectrum?

j. What are the advantages of a grating spectrograph as compared with a prism spectrograph? What are the disadvantages?

k. What are the advantages of spectral analysis as compared with chemical analysis?

l. Can you suggest applications of spectroscopy to (1) medicine? (2) biology? (3) criminology? (4) astonomy?

m. Give examples of a source of a (1) continuous, (2) line, (3) band spectrum.

n. If you were to start moving with a speed of 10^{10} cm/sec toward a source of red light, what would happen to the color which you perceive?

o. How does the Doppler effect depend on the distance between the source and the observer?

p. If you were to study a double star by means of a spectroscope, what would you see?

q. Could you detect rotation of a planet by means of the Doppler effect?

r. How can you determine the chemical composition of granules contained in a living cell without killing or injuring the cell?

Spectroscopy

We have seen that the different colors of light correspond to different frequencies of ether vibrations just as differences in pitch of musical sounds are associated with differences in the frequency of vibration of air particles. In distinguishing between two colors, our eye is differentiating between two frequencies of vibration just as an ear is when it differentiates between two sounds of different pitch. But in one respect our ear is much superior to our eye. When an instrument in a practicing orchestra plays a wrong note, a good conductor is usually able to determine what note it was and what instrument was the guilty party. If we imagine what a complicated curve the oscillogram of a musical chord played by an orchestra is and furthermore remember that the tone characteristics of an instrument are determined by the relative intensities of various overtones, we realize what a veritable miracle of harmonic analysis is performed by the auditory sense of a skillful conductor in separating in his mind the sounds produced simultaneously by different instruments. If, on the other hand, we mix three or four different paints together and ask the most skillful painter to tell us the color components of the mixture, we shall see from his dismal failure how much inferior our eye is to our ear as a harmonic analyzer.

Those who like teleological reasoning may argue that nature did not equip our eye with analytical ability comparable in subtlety to our ear because it would be of little use. But this is a wrong idea. As we shall see later, the development of special instruments which can be considered as "optical harmonic analyzers" has led to spectacular advances of our knowledge of nature and our mastery over our environment. Instruments which analyze colored light into the components of the color mixture and enable us to estimate the relative content of light of different frequencies in the mixture are called **spectroscopes.**

The most obvious application of spectroscopes would be to the study of matter, *i.e.*, to chemistry. One of the most important distinguishing features between different substances is their color. Some materials absorb red light out of a bundle of illuminating white light in preference to other colors, and the substance appears green. In other cases there may be a different pattern of absorption and the substance will appear in a different color whose "shade" may not be easy to describe in words.

The Grating Spectroscope. A quantitative way of describing color can be based on the following procedure: Send the beam of light, which is to be analyzed, through a diffraction grating. Light of different frequencies will be diffracted through different angles. We can infer the wave length of the light moving in a particular direction from the corresponding deviation. A thermocouple could be exposed to the radiations of different color, which have been thus separated, in order to determine how the total energy of the mixed light beam is distributed among the different colors.

Figure 31.1 shows a grating spectroscope which utilizes the principle described above. L is a source whose light is focused on the narrow slit S (the slit is perpendicular to the plane of the paper) by the lens L_1. We consider this slit now as a secondary source of light and as the object of the lens C. Since S is placed in the focal plane of C, light rays emanating from every point of the slit are rendered parallel by C and pass through the diffraction grating D. Let us assume that our source is a monochromatic one emitting, say, red light of a definite wave length. Under these conditions, in addition to the direct (*i.e.*, undeviated) transmitted beam *O*, we shall detect diffracted beams *R* and *R′* which correspond to the first order of diffraction. There may be more strongly deviated red beams due to a higher order of diffraction, but we shall limit ourselves to the diffraction spectrum of the first order, which is the most intense.

A telescope *T* is mounted on the circular graduated table so that it can be rotated about the center of the table pointing toward the grating. If we move the telescope so as to intercept beam *O*, we see the red image of the slit in the center of our telescopic field of view which is marked by a cross hair. If we move the telescope to positions where the beams *R* or *R′* are intercepted by it, we see again a much fainter red image of the slit. If we rotate the telescope through larger angles, we may encounter still weaker red slit images which are due to diffraction of higher order.

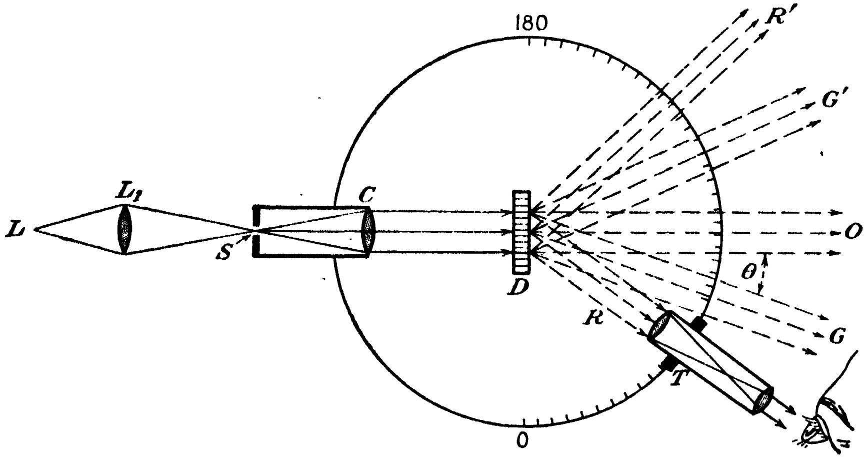


Fig. 31.1. A grating spectroscope.

Suppose now that instead of using a source of red light, we used a source *L* of pure green light. In this case we should have seen the same thing except that our slit images would have been green and the angles of deviation of the diffracted beams would have been smaller. (Rays *G* and *G′* are due to first-order diffraction.) In all cases the angle of deviation can be determined exactly by reading the telescope position on the graduation of the circular table.

Let us now assume that we have a source which emits simultaneously blue-green and red light. Since these colors are complementary, the source will appear white. When we view the direct beam *O*, we see a white slit image, but when we swing the telescope into new positions, we see at first a green slit image as we intercept ray *G* and then a red one when we intercept R. (If higher orders of diffraction were present, we would encounter alternately additional green and red images.)

If we limit ourselves to the spectrum of the first order and observe only on one side of the direct beam, we can say quite generally that we see as many distinct slit images as there are distinct colors of light in our light mixture. Sources emitting several discrete colors which give rise to several slit images of different color can be obtained by passing an electric current through an atmosphere of a rarefied gas (gas-discharge tube).

Equation (29.3) enables us to determine the wave lengths corresponding to each of the colors emitted by measuring the angle θ which the diffracted beams form with the direct beam O. By selecting a source which emits many colors between red and violet, we can plot a graph of wave length versus angle θ and thus *calibrate our spectroscope* so that, if we should see a first-order slit image with an unknown light source at the telescope setting of $\theta = 29.56°$, we shall know that light of, say, $\lambda = 5{,}012$ A is emitted by the source.

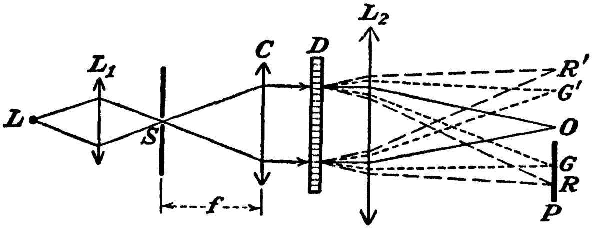


FIG. 31.2. Scheme of a grating spectrograph.

The individual slit images are called **spectral lines.** If, instead of an illuminated slit, we used an illuminated circle or a triangle as our secondary source of light S, we should be able to speak of spectral circles or spectral triangles instead of spectral lines.

Instead of visual observation, we can also resort to photographic recording of the spectral lines. In this case we call the instrument in which a fixed camera replaces the movable telescope a **spectrograph.** Figure 31.2 shows a self-explanatory diagram of a grating spectrograph. L_2 is the lens of the camera, and P is the photographic plate. G and R are the recorded spectral lines, *i.e.*, slit images revealing the presence of red and green light in the source L. O, G', and R' are the slit images formed by direct (O) and diffracted rays (G' and R') which are not photographed. S is the slit, C is a lens which renders the light coming from S parallel, and D is the diffraction grating.

The Prism Spectrograph. A prism could also be used as a spectrographic device. In fact, the first spectroscopes utilized a prism rather than a grating to analyze mixed light into its color components. Figure 31.3 shows a simple prism spectrograph.

The light of the source *L* is concentrated on the slit *S* which lies in the focal plane of the lens *C*. Conical bundles of light issuing from *S* are rendered parallel by the lens *C* and are subsequently refracted by the prism *Pr*. The lens L_2 focuses the light issuing from the prism upon the photographic plate *P*.

Since there is no way of computing the wave length of light from its deviation by a prism, spectrographs of this type are calibrated by photographing a known line spectrum in which the emitted wave lengths have been determined by means of a diffraction grating. This enables one to draw a calibration curve by plotting wave length versus deviation.

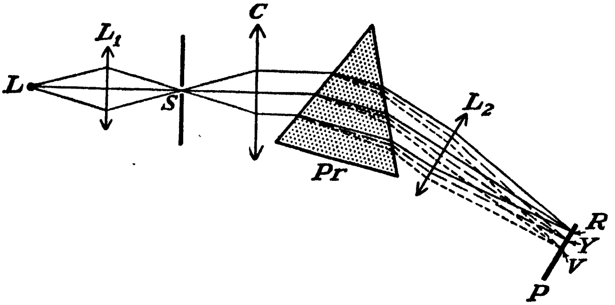


Fig. 31.3. Scheme cf a prism spectrograph.

The advantage of the prism spectrograph over the grating device lies in the high intensity of the slit images produced by it. The whole of the light energy of one color is concentrated in one image rather than a fraction of it as in the case of the first-order diffraction image. The prism spectrograph is useful in cases where weak sources of light are to be studied, *e.g.*, faint stars.

The grating spectrograph, on the other hand, can separate light of different colors much more widely and is more useful than the prism instrument where the study of fine spectral detail is required.

Types of Spectra

In using a spectroscope to analyze the color of chemical compounds and elements for the purpose of identification, we can proceed in different ways. We can, for instance, make the substance under investigation emit light and use this light as source *L* in the setup of Fig. 31.2 or 31.3. This can be done, for instance, in the case of certain light metals by heating a sample of a metal salt in a bunsen flame or in the case of a gas at low pressure by passing an electric current through it by applying a high voltage across the gas-discharge tube. We obtain in these cases so-called **emission spectra,** which are designated more specifically as emission **line spectra** when the photograph taken with the setup of Fig. 31.2 or 31.3 shows a series of thin lines (slit images), indicating the presence of light of several discrete colors or, more specifically, of several discrete frequencies (or wave lengths).

Not all luminous sources give a line spectrum. If, for instance, the gas pressure in the gas-discharge tube is increased, the lines in its spectrum become wider. At very high pressure the spectral lines finally become so wide that neighboring lines fuse together, creating a continuous sequence of changing color (**continuous spectrum**). Any glowing solid or liquid object (provided it is not surrounded by an appreciable atmosphere of vapor) emits a continuous spectrum (for instance, the incandescent filament of a tungsten lamp). The continuous visible spectrum of a white source extends from red to violet.

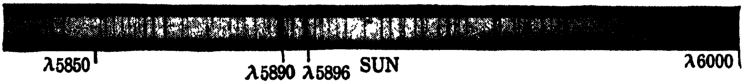


Fig. 31.4. The Fraunhofer absorption lines in the sun's continuous spectrum. (*From Jenkins and White, "Fundamentals of Optics," McGraw-Hill.*)

If we now send white light of a source which emits a continuous spectrum through various colored solutions or through some gases and vapors, we notice that the transmitted light is not white any more. For instance, white light sent through a copper sulfate solution emerges as blue light. This suggests that some of the red light originally contained in the white radiation has been absorbed. In the case of gases and vapors the continuous spectrum is frequently traversed by black lines, *i.e.*, discontinuities in the spectrum, which indicate that light has been absorbed in certain narrow regions of frequency. Figure 31.4 shows a **line absorption spectrum.**† Colored solids and liquids give **continuous absorption spectra** in which wide regions of frequency are eliminated. Figure 31.5 shows absorption spectra of various colored glasses.

One and the same element may be used to give both an absorption spectrum and an emission spectrum. For instance, when sodium vapor is heated to a high temperature, a yellow line appears in its emission spectrum. If, on the other hand, we use sodium vapor of relatively low temperature and send a beam of white light through it, we obtain a black line in the yellow region of the emission spectrum. In fact, *the black absorption line appears at exactly the same place as the yellow emission*

---

† Solar spectrum in the neighborhood of the Fraunhofer *D* lines. The two strong lines are due to absorption by sodium vapor in the solar atmosphere.

line, which means that the sodium vapor is capable of absorbing light of exactly the same frequency as it emits.

Extended investigation of this phenomenon with different gases showed that in general *all the lines appearing in the absorption spectrum can be found in the emission spectrum* but not vice versa, since the emission spectrum usually contains more lines. The reason for this difference between absorption and emission spectra will be discussed in Chap. 36.

This relation between absorption and emission spectra can hardly be fortuitous. We shall offer a tentative plausible explanation for it which, as we shall see in Chap. 36 on atomic theory, has to be modified radically. Imagine a singer to be separated from his audience by a frame which carries many rows of densely spaced and well-tuned piano strings. The first row contains only *a* strings, and the second row only *c* strings. Any of these strings will vibrate and emit the tone *a* or *c* when plucked. Now as the singer sings, the *a* and *c* strings will resonate whenever he happens to sing *a* or *c*. Not all the energy absorbed by the strings, however, will be reemitted in the form of sound of the same frequency. There will be other processes, such as emission of sound of other frequencies and degradation of elastic sound vibration to disorderly heat motion (through friction), which will account for some energy apparently being "lost" or *absorbed* by the intervening "medium." Similarly, we may imagine a gas to contain atomic resonators which selectively absorb light of certain frequencies only, namely, the frequencies which they are known to emit in emission spectrum. Since they only reemit part of the absorbed energy in its original form, there will be a relative weakening of intensity just at those frequencies when white light passes through such an atmosphere.

FIG. 31.5. Absorption spectra of various colored glasses. The top spectrum is the continuous spectrum of the light sent through the glass. (*From Jenkins and White, "Fundamentals of Optics," McGraw-Hill.*)

In addition to line spectra, which are observed in emission as well as absorption spectra, one also encounters so-called **band spectra** in emission and absorption. Figure 31.6 shows the photograph of a band spectrum.

A study of a band spectrum with a high-resolution spectroscope reveals a very fine line structure which creates the illusion of shading. Simple line spectra (as distinct from band spectra) are atomic spectra. They are emitted by gases in which some atoms are present in the free state, whereas band spectra are molecular spectra.

A survey of spectra of various compounds and elements revealed that the spectrum of a substance is a very specific characteristic of the substance which can be used to identify all species of matter. The study of the spectral "fingerprints" of matter has established a most powerful means of chemical analysis. The sensitivity of spectral analysis exceeds by far the sensitivity of chemical detection. For instance, 10^{-10} gm of lithium can be detected spectroscopically. One has merely to look for the characteristic lithium lines, whose position in the spectrum is known, among the lines which are due to other components of the analyzed mixture. Chemical analysis utilizes both emission and absorption spectra.

FIG. 31.6. A band spectrum. (*From plates by Prof. J. F. Jenkins.*)

Chemical Analysis of Stars and Planets

Perhaps the most impressive applications of spectral analysis are those made in astronomy. One might say that the development of the science of **astrophysics** is largely due to the application of the spectrograph to the study of stars. The spectrograph enables us to perform a chemical analysis of specimens of matter which are so far from our laboratories that light has to travel millions of years to cover that distance. In one instance, in fact, a chemical element was discovered in the atmosphere of a star by means of spectroscopy before its existence on earth was detected. It was helium (from Greek *helios,* meaning the sun) whose presence in the sun's atmosphere was revealed by unfamiliar absorption lines in the sun's spectrum which were attributed to a new element. This element was later shown to be present in our own atmosphere in small concentrations.

The composition of the sun's atmosphere is revealed by numerous absorption lines which interrupt the continuity of the sun spectrum (Fraunhofer lines, see Fig. 31.4 which shows only a small section of the sun's spectrum in the vicinity of the *D* lines). All these lines can be shown to correspond precisely to wave lengths emitted by known terrestrial elements, and so the presence of hydrogen, calcium, sodium, and many other familiar elements can be inferred from the study of the Fraunhofer lines. The solar spectrum suggests that the interior of the sun is dense enough to give rise to a continuous emission spectrum and that this dense core is surrounded by a rarefied gaseous atmosphere which is largely responsible for the absorption spectrum. We should naturally expect some of the absorption lines to be due to absorption of light in the earth's atmosphere. Similarly, we shall expect the atmospheres of the planets to absorb specific frequencies of light peculiar to the spectrum of the elements and compounds which occur in their gaseous envelopes. Planets shine by reflecting the sunlight. Hence, light coming from a planet has passed in addition to the sun's and earth's atmosphere twice through the atmosphere of the planet, and indeed, we discover absorption lines in planetary spectra which indicate the presence of compounds in their atmosphere which are absent in the earth's and sun's atmosphere. Thus, for instance, ammonia vapors and methane gas have been detected in the atmosphere of Jupiter. The moon has no atmosphere. The absorption spectrum of moonlight is the same as that of sunlight. Similarly, the chemical composition of the atmosphere of distant stars and island universes (spiral nebulae) can be determined.

Spectroscopic Thermometry

But the spectroscope is not merely the astronomer's test tube; it also serves him as a thermometer. If, for instance, we gradually heat a solid object (say, a platinum wire), it looks at first red, then yellow and finally becomes white. It is more revealing to watch this change in appearance through a spectroscope. We see, then, how at first only the red end of the spectrum appears, then at a temperature of about 1200°C orange and yellow make their appearance in addition to red. At about 1700°C the glowing wire looks yellow and its spectrum reveals an appreciable emission of green light in addition to red, orange, and yellow. At about 2700°C the glowing body appears white—all spectral colors appear in the spectroscope. Further temperature rise increases the brightness of the body but causes no change in its color. All substances behave alike in the performance of this experiment. The continuous spectrum emitted by a solid body is not characteristically determined by its composition. Only line or band spectra (emitted or absorbed by gases or absorbed by solutions) are characteristic of the substance.

The preceding experiment suggests the possibility of inferring the temperature of a body from a study of its continuous emission spectrum. It is more appropriate for a quantitative study to use a sensitive thermometer as a detector of radiation rather than our eye. Figure 31.7 shows a scheme in which the radiation of the hot body B passes through the prism P which resolves it into a spectrum. The system of lenses which focuses the spectrum in the plane Pl is omitted for simplicity. A thermometer T can be placed in any portion of the resulting spectrum. R is its red end, and V the violet end. By thermometric measurements in different regions of the spectrum a comparative measure of the energy radiated in the different spectral domains can be obtained. Figure 31.8 shows a series of such energy-distribution curves obtained with the body B at different temperatures.

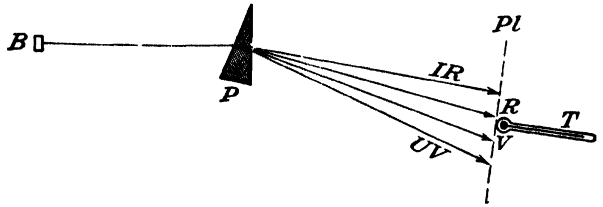


FIG. 31.7. Scheme of a thermometric spectrometer.

We see how the maximum of the distribution curve shifts toward shorter wave lengths as the temperature of the body increases. Wien found a simple empirical relationship between the location of the energy maximum and temperature, which is known as **Wien's displacement law:**

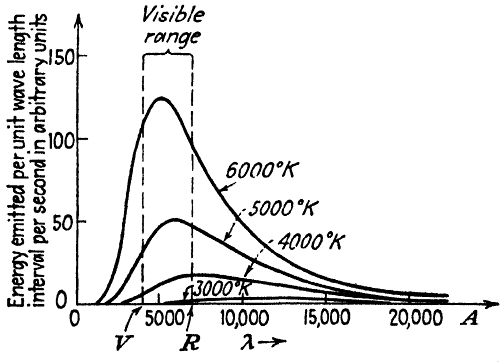


FIG. 31.8. Spectral distribution of energy emission of a black body at different temperatures. (*After Jenkins and White, "Fundamentals of Optics," McGraw-Hill.*)

$$\boxed{\lambda_{\max} = \frac{C}{T}} \tag{31.1}$$

$\lambda_{\max}$ is the wave length at which the energy maximum is located. T is the absolute temperature of the body, and $C = 0.2897$ cm deg is a constant. We see that doubling the absolute temperature halves the wave length at which the maximum emission occurs. The spectroscope enables us to find $\lambda_{\max}$ and thus to determine the temperature of the hot body.

The knowledge of the star's temperature (which can be determined with the aid of Wien's displacement law) in conjunction with the knowledge of the illumination produced by the star on the earth's surface enables us to compute the diameter of the star with the aid of Stefan's law of radiation (see Chap. 17). Thus, spectroscopy also provides a yardstick, as it were, to measure stellar diameters.

Spectroscopic Speedometry

But these applications do not exhaust the versatility of the spectroscope. In the hands of astrophysicists this instrument was turned into a reliable speedometer for stars. In Chap. 27 we showed how the pitch of a sound may be varied by relative motion of the source and the observer (**Doppler effect**). A speedometer enabling us to measure the relative velocity between the source and observer along the line joining the two can be based on the Doppler effect. This effect was shown to be a general phenomenon which should be observable with any type of wave motion. Thus, just as an observer hears a higher pitch when moving toward the source of sound, he should, when moving toward a light source, see light of a higher frequency than he would have seen when at rest relative to the source. In other words, a red lantern should look blue to him if he were moving fast enough toward it or if it were moving equally fast toward him. Or conversely, a blue lantern should look red to the observer if the distance between him and the lantern were to increase fast enough.

In the case of light the two formulas (27.6) and (27.7) can, owing to the high value of the velocity of light c, be replaced by one single approximate formula which applies to both cases:

Replacing V by c in (27.6) we can write for the case of approach

$$P'_a = f\left(\frac{c}{c-s}\right) = f\left(\frac{1}{1-s/c}\right) \qquad (a)$$

and for (27.7)

$$P'' = f\left(\frac{c+b}{c}\right) \qquad (b)$$

We can write

$$\frac{1}{1-s/c} = \left(\frac{1+s/c}{1+s/c}\right)\frac{1}{1-s/c} = \frac{1+s/c}{1-(s/c)^2} \qquad (c)$$

Since in most practical cases the velocity of the source will be much smaller than the velocity of light ($s \ll c$), the square of the ratio s/c will be negligible as compared with 1.

Hence, we can write for equation (c)

$$\frac{1}{1 - s/c} \cong 1 + \frac{s}{c} = \frac{c}{c} + \frac{s}{c} = \frac{c + s}{c} \qquad (d)$$

Substituting (d) in (a) we get

$$P'_a = f\left(\frac{c + s}{c}\right)$$

This is the same as equation (b) except that s stands in the place of b (where b is the velocity of the observer).

Thus we can use one single *Doppler formula*† *for light:*

$$\boxed{P = f\left(\frac{c + v}{c}\right)} \qquad (31.2)$$

where v is the relative velocity (of approach) between source and observer regardless of who is moving relative to the other, f is the frequency of the light as determined by an observer at rest with respect to the source, and P the frequency as determined by the observer who is moving with respect to the source.

How could we check in our laboratory whether or not there is a Doppler effect in the case of light? It would be very difficult to move a source of light with a speed large enough in comparison with c to get an appreciable Doppler shift of color! Fortunately, we are in possession of minute sources of light which can be moved with speeds comparable to c without much difficulty. We possess in canal rays (positively charged ions accelerated to high speeds) minute fast-moving sources of light. We should expect the lines in their emission spectrum to be shifted toward the blue when we observe the miniature sources of light as they move toward us and toward the red as they move away from us. Furthermore, knowing the accelerating voltage and the mass and charge of the ions, we know their speed and can therefore predict the magnitude of the expected Doppler shift. Spectroscopic observation should allow us to check our predictions in a quantitative manner. Experiments of this kind have been performed and have verified the existence of the Doppler effect with light.

---

† This formula applies to the case of mutual approach of observer and light source. When the two recede from each other replace the plus sign by a minus sign.

Doppler effect explains why spectral lines in the emission spectrum of a rarefied gas become widened as the gas is heated.† Suppose we observe the emission spectrum at a very low temperature at which the atoms move with negligible speed and observe a spectral line of a definite width, *i.e.*, of a wave-length range extending from, say, $\lambda' = 4{,}000.0001$ A to 4,000.0002 A. When we heat the gas, we observe at a certain temperature that the range of λ has been extended to the interval of 4,000.0000 A to 4,000.0003 A. This "widening" of the spectral line can be attributed to a Doppler effect. The atoms of the gas move faster at the higher temperature. Many of the atoms emit light as they move toward us, and we see light of a shorter wave length while others move away from us, emitting light which appears to us to have a longer wave length than light emitted by atoms at rest.

The successful demonstration of the Doppler effect with light can be accepted as additional evidence for the wave nature of light.

Astronomers realized that the Doppler effect offers a means of measuring the "radial velocity" of stars (*i.e.*, the component of the relative velocity with respect to us in the direction of the line of sight) by means of the spectrograph. The absorption lines of certain elements in the spectrum of a star which is moving away from us should not be exactly in the place (*i.e.*, at that frequency) where they appear in absorption spectra taken with our laboratory sources. They should be shifted from their normal position toward the red end of the spectrum. The magnitude of the observed shift should permit the determination of the radial velocity of the star. Spectroscopic measurements of stellar velocities proved eminently successful. It was found that the stars are moving in all directions with various speeds, some of them approaching us with velocities as high as 54 km/sec. A particularly interesting discovery was made in connection with extragalactic nebulae. Some of the *neighboring* galaxies (*e.g.*, the Andromeda nebula) are moving toward our own galaxy, and others move away from it. But the *distant ones* all show a marked red shift in their spectra which, if interpreted as a Doppler shift,‡ indicates that they recede from our galaxy. The speed of recession turns out to be directly proportional to the distance of a nebula from us. Some of these gigantic systems have been observed to flee from us with about one-third of the velocity of light. Hubble proposed the hypothesis that the observed recession of spiral nebulae is due to an expansion of the universe (compare Chap. 40).

---

† A complete explanation of the broadening of spectral lines takes into account, in addition to the Doppler effect, other significant effects.

‡ The Doppler shift hypothesis is not the only possible interpretation of this red shift, but it is at present the most widely accepted one.

The spectroscope has helped astronomers to solve many puzzles and has presented them in addition with new riddles. For instance, there are some stars whose spectrum shows the following peculiar behavior: On a certain day we may see an absorption spectrum of normal appearance, but as we watch the spectrum for several days or weeks, we see how each spectral line splits into two lines. The new lines move away from the original position of the line, moving in opposite directions and usually at unequal rates. After a while the lines begin to converge toward each other until they meet and fuse into one at the location of the original line, whereupon the process is repeated, with the "slow" line which previously moved to the right now moving to the left. This phenomenon is immediately explained when we point a telescope at a double star which can be resolved into two separate stars in a telescope. Figure 31.9 shows two nearly equal stars revolving about a common mass center C and the telescope T through which the stars are observed. When the stars are in positions A and A', the absorption lines appear in their normal positions, but when the stars are in positions B and B', the deviation of the lines is at a maximum. It is easy to understand why it is so. The radial velocities of the stars with respect to the observer are greatest for this configuration, so that the spectrum of the star at B' displays a maximum Doppler shift toward the blue while the star at B exhibits a maximum red shift. As the stars move from position A to position B', the orbital velocity component aiming toward the observer varies and so does the position of the absorption lines in the stars' spectrum. In general the two stars have unequal mass, hence unequal orbital velocities with respect to a mass center which does not bisect a line joining their centers. This explains the asymmetry of the observed Doppler effect, *i.e.*, the "fast" and "slow" lines referred to above.

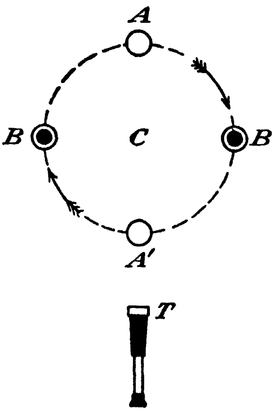


FIG. 31.9.

Only few binary systems can be resolved telescopically into two separate stars. In most cases we can only infer from the observed periodic shift of spectral lines that what appears in the telescope as a single star is actually a double star. Double stars are not at all uncommon. About 25 per cent of all the stars are double stars. Triple, quadruple, and even quintuple stars have also been observed. Not only stars but even whole galaxies can form binary systems, or double galaxies, which revolve about a common mass center. This, too, has been revealed by observation of the Doppler shift of spectral lines.

The Doppler effect can also be used to detect rotation of a star, a planet, or a galaxy and to measure the angular speed of rotation. Imagine that the dotted circle in Fig. 31.9, instead of representing the orbit of two stars, represents the cross section of the sun. Assume the sun to rotate about its axis as shown by the arrows. If we examine separately the spectra from the edge B, edge B', and the center A' of the sun's disk, we should find the following: (1) a Doppler shift toward the blue for edge B', (2) a red shift for edge B, and (3) a normal spectrum for center A'. This is, indeed, what we observe, and the amount of the Doppler shift gives the correct period of rotation of the sun of about 27 days, which has been found previously by observation of the sunspots.

By the same method it was shown that Saturn's ring is rotating and that the innermost parts of the ring revolve faster than the outer parts, which proves that it cannot be a solid ring but rather must consist of debris of matter which circles about Saturn like a cloud of satellites.

The analysis of the chemical composition, dimension, and motion of distant celestial bodies constitutes undoubtedly one of the greatest scientific achievements of the human mind.

Invisible Light

Let us suppose that we have a source of green light. If it moves fast enough toward us, it may appear violet and have a wave length of $\lambda_V = 4{,}000$ A; if it moves fast enough away from us, it may appear red and have a wave length $\lambda_R = 8{,}000$ A. What color will it have if it moves so fast toward us that its wave length becomes shorter than 4,000 A or when it recedes from us so rapidly that $\lambda > 8{,}000$ A? This question suggests that there must exist ether vibrations of higher and lower frequency than those of the familiar colors of light. We might then expect to discover by this experiment yet unknown colors or to find that our eye is insensitive to radiations of those new frequency ranges just as our ear does not perceive sound of frequencies below 16 or above about 20,000 cycles/sec. If the latter is the case, then it is possible that such ether waves emanate from familiar sources of radiation but that they cannot be detected unless we replace our eyes by a detector which will respond to them.

This is actually the case. If we use a quartz prism instead of a glass prism in the experiment shown in Fig. 31.7, the thermometer indicates the presence of less refrangible radiations beyond the red region (between ray R and ray IR), which we call **infrared radiations,** and of a more refrangible invisible radiation beyond the violet region (between ray V and ray UV) which we call the **ultraviolet radiation.** The curves of Fig. 31.8 show that at low temperatures most of the energy of the hot radiator is issuing in the form of infrared radiation ($\lambda > 8{,}000$ A). As the temperature increases, the emission in the ultraviolet region is seen to go up. The energy emitted in the form of visible radiation constitutes a small portion of the total emitted energy in all of the cases.

Hot bodies are a common and efficient source of infrared radiation, which is also referred to as *heat radiation.* It is possible to perform the same diffraction experiments with heat rays as with light, which prove their wave nature and enable us to measure their wave length. Since during the solar eclipse heat rays are cut off at the same time as light, we know that they travel in a vacuum with the same speed as visible light. Measurements made with diffraction gratings show that heat rays have wave lengths longer than 8,000 A. The near infrared is transmitted well through glass, but for wave lengths beyond 20,000 A rock-salt prisms and lenses must be used to avoid absorption in glass.

Up to $\lambda \simeq 13{,}000$ A infrared radiations can be detected by photographic plates. Infrared photography is of great practical value, since fog does not scatter infrared rays nearly so much as light, and thus it is possible to take photographs through fog using an infrared film in conjunction with an "infrared filter." A solution of iodine in CS_2 or alcohol is opaque to visible light but transparent to infrared, and hence, it can be used as an infrared filter. The infrared rays penetrate more deeply through the human skin than light, so that infrared photography is also useful in diagnosing, for instance, varicose veins, superficial tumors, and thrombosis of superficial vessels.

A most sensitive general method of detecting infrared radiation is by means of bolometers or thermocouples. By such means astronomers have been able to detect dark stars which are so cool that only infrared radiation is emitted by them in appreciable amounts.

The *ultraviolet radiation* was discovered by Ritter through its photochemical effectiveness. He allowed a sun spectrum to be formed on a piece of filter paper soaked in silver chloride and noticed that it was blackened beyond the violet end of the visible spectrum.

As in the case of infrared light the wave nature of ultraviolet radiation can be demonstrated with diffraction gratings. Their wave length turns out to be shorter than the wave length of violet light, which is 4,000 A.

As the ultraviolet rays traverse the air, an odor is noticed which is due to formation of NO_2 from the nitrogen and oxygen of the air. At the same time a gas is formed called "ozone" (the "smelling gas"), which owes its name to the fact that the odor of NO_2 was originally mistakenly ascribed to it. An ozone molecule consists of three oxygen atoms. It absorbs ultraviolet rays of a certain range of wave length very strongly. This is a fortunate circumstance. The short waves of ultraviolet are very harmful to living organisms. (They are successfully used to sterilize the air in operating rooms.) The fact that these rays cause ozone formation in the atmosphere and that this gas absorbs ultraviolet strongly in the upper strata of the atmosphere saves the organisms on the earth's surface from exposure to these lethal rays. Figure 31.10 shows the lethal effect of ultraviolet light upon paramecia. Figure 31.10*a* shows normal protozoa; Figure 31.10*b* shows the microorganism swollen after 30 sec of irradiation. Figure 31.10*c* shows beginning disintegration of the animal, and Fig. 31.10*d* shows how the unicellular organism bursts so that its cell content is mixed with the surrounding fluid.

The pigmentation of the human skin by ultraviolet radiations (ery-

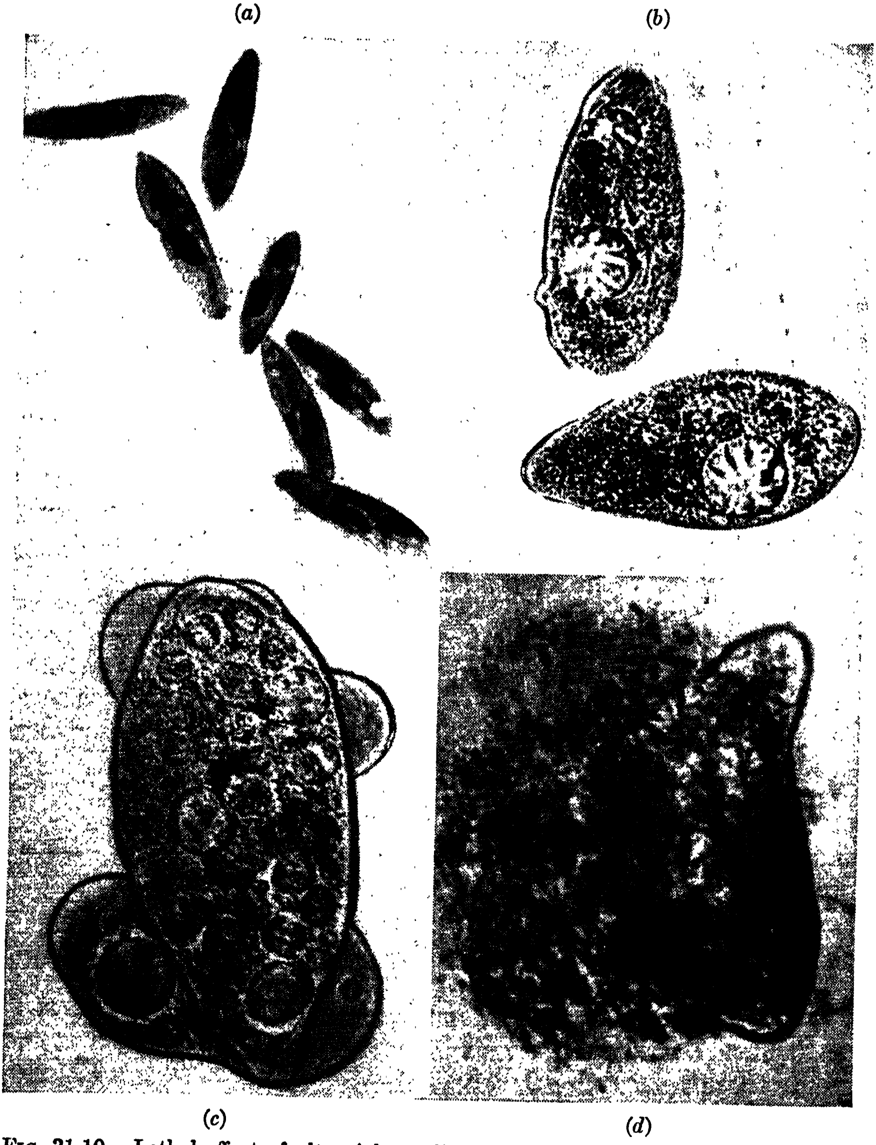


FIG. 31.10. Lethal effect of ultraviolet radiation upon paramecia. (*Courtesy of the Westinghouse Electric Corporation.*)

thema) is a well-known biological reaction. Ultraviolet irradiation of the skin promotes the formation of vitamin D from cholesterol.

Fluorescence and Phosphorescence

Ultraviolet light can be detected by photographic plates and by the nonspecific method of thermometry (thermocouples or bolometer). In many applications the most convenient way of detecting ultraviolet radiation makes use of its ability to excite fluorescence. Many substances are made to emit visible radiation when illuminated by ultraviolet light. This is not a specific property of ultraviolet rays. Fluorescence may be excited by visible light too. For instance, if a *thin beam* of yellow light derived from a source of glowing sodium vapor is sent through a large vessel containing sodium vapor, the whole vessel lights up in the typical color of the yellow sodium emission line. The sodium vapor reemits light of exactly the same frequency as the light which is used to excite the fluorescence (*resonance fluorescence*). This is, however, not a typical case. **Usually the reradiated fluorescent light is of longer wave length than the illuminating light (Stokes' law of fluorescence).**

For instance, nearly all petroleum products, such as vaseline, kerosene, and heavy oils, glow with a bluish color when exposed to white or to violet or ultraviolet light but do not glow when exposed to green light. The conversion of ultraviolet into visible fluorescent radiation is in accordance with Stokes' law. Very many substances show typical fluorescent colors when irradiated by ultraviolet light. This effect has found numerous applications in many fields. For instance, almost any body fluid can be used as an invisible ink. When the letter which appears blank in white light is exposed to ultraviolet, the message shows up in brightly glowing letters. In many jails outgoing mail is inspected under ultraviolet illumination.

Fluorescence can also be used to detect forgeries, since erasures change the character of the fluorescence of the paper and different inks fluoresce with different colors when subjected to ultraviolet irradiation.

Bloodstains can be easily distinguished from paint owing to difference in the color of fluorescence. False teeth can be distinguished from living teeth, since the latter fluoresce strongly when exposed to ultraviolet light whereas the former do not.

Various carcinogenic substances (cancer-producing chemicals) can be identified by their fluorescence.

Some methods of quantitative determination of vitamins and enzymes are based on the measurement of the intensity of fluorescent radiation.

Various skin diseases can be diagnosed by exposing the skin to ultraviolet light. For instance, ringworm (a fungus infection of the skin) causes intense yellow or green fluorescence.

Ultraviolet light can be used in conjunction with quartz lenses in specially designed microscopes. The use of ultraviolet light lowers the wave length and, hence, allows one to obtain a higher resolution and thus to use higher magnification.

In studies of fluorescence caused by ultraviolet light, a quartz mercury arc lamp is commonly used as a source of ultraviolet radiation. Special dark glass (Wood's glass) is used to filter out visible light, transmitting almost only the ultraviolet.

A body is called fluorescent if the visible radiation vanishes instantly as soon as the ultraviolet is turned off. There are substances, however, which continue emitting visible light for an appreciable time after exposure to ultraviolet radiation (or visible radiation of short wave length). This phenomenon is called **phosphorescence,** and substances which display it are called *phosphors.*† For instance, the mineral zinc blende (ZnS) emits phosphorescent green light after exposure to blue, violet, or ultraviolet light.

If a phosphorescent screen which emits light is illuminated by infrared light or heated, the phosphorescence is quenched and the screen becomes dark at the spot exposed to heat rays. The rate of emission of phosphorescent radiation depends on the temperature of the screen. When the screen is cooled appreciably, its brightness diminishes but it continues emitting light for a longer period of time. A screen may be cooled to such a low temperature that the phosphorescent emission becomes imperceptible. If such a screen is exposed to infrared radiation, the rate of phosphorescent emission is speeded up at the irradiated spot, which is thus made visible by a bright glow. This effect makes it possible to convert infrared into visible light, but the necessity of cooling the screen makes the method cumbersome. Recently, screens have been developed which allow us to accomplish the same result without cooling the screen. The screen is exposed to ultraviolet light but emits very little light thereafter. Spots which are subsequently hit by infrared rays cause emission of green light. The "snooperscope" and "sniperscope" utilize such screens. The victim is illuminated with invisible infrared rays. A special camera creates an infrared image on the screen of the type described above. The screen converts it into a visible image and allows the sniper to take his aim.

Such screens may find peacetime uses in devices enabling us to see through fog by means of infrared radiation and in biological observations by means of infrared light.

---

† They have nothing to do with phosphorus, which is not phosphorescent. It emits light owing to slow oxidation of phosphorus to P_2O_3.

Spectroscopy in Biology

Chemical methods of detecting the presence of biologically important substances, such as vitamins, hormones, and enzymes, are usually very cumbersome and time-consuming. This is even more true of the methods of biochemical quantitative analysis. The possibility of using a spectroscope for qualitative and quantitative biochemical analysis is of tremendous practical value.

Thus, for instance, the presence of hemoglobin in body fluids can be recognized by the characteristic absorption bands of hemoglobin in the visible region. Similarly its oxidation product, the oxyhemoglobin, and the stable compound between carbon monoxide and hemoglobin which is formed in cases of illuminating-gas poisoning can be recognized by their band absorption spectrum in the visible range. In some cases the colorless substance which is to be detected is allowed to react with a suitable reagent (such as ninhydrin in the case of proteins) as a result of which a colored product is formed which shows a characteristic absorption spectrum.

Quantitative spectral analysis is performed with so-called *spectrophotometers*, which are a combination of a photometer and a spectroscope. The latter produces a spectrum from which a narrow portion is allowed to enter through a slit into the photometer, which measures the intensity of the transmitted light. By comparison of the light transmission in a selected region of the absorption spectrum through a standard solution of known concentration with the solution of unknown concentration, the concentration of the latter can be determined.

In cases of colorless solutions one can frequently use spectroscopic and spectrophotometric methods by utilizing the absorption bands in the ultraviolet region.

In studying the interior of unicellular organisms and of individual cells of various tissues, biologists discovered tiny particles to which a variety of names has been given, such as centrioles, nucleolus, and chromosomes, which are not indicative of their chemical nature. It is of the greatest biological interest to determine what these particles consist of and how their substance changes in the course of biological processes such as cell division. At this point the reflecting microscope enters the picture in its capacity as a tool for spectroscopic microanalysis which is the revolutionary feature of this instrument alluded to in the preceding chapter. Its achromatic features make it possible to transilluminate a cell with visible as well as ultraviolet and infrared light while a portion of the enlarged image of the cell remains focused on the slit of a spectrometer. Thus we can get the entire absorption spectrum of a small portion of the cell (for instance, of a small intracellular object) in a single exposure and identify its chemical composition. In particular, the use of infrared light looks very promising because of the characteristic nature of the infrared absorption bands of organic molecules, which have been labeled by some as the "fingerprints of the molecules." With the aid of the reflecting microscope spectral analysis of crystals of a mass less than 10^{-10} gm has been performed.

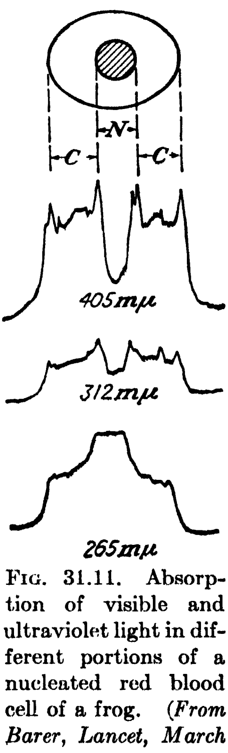


FIG. 31.11. Absorption of visible and ultraviolet light in different portions of a nucleated red blood cell of a frog. (*From Barer, Lancet, March* 26, 1949.)

Figure 31.11 shows as an example of microscopic spectral analysis the absorption of visible and ultraviolet light of selected wave lengths in different regions of the nucleated red blood cell of the frog. The ordinates of the curves shown under the picture of the blood cell with its nucleus represent the per cent absorption of a thin beam of light which we can imagine to wander across the cell. We see that at $\lambda = 405$ mμ (= 4,050 A) there is a strong absorption in the cytoplasm and very little absorption in the nucleus. At $\lambda = 3{,}120$ A there is little variation in absorption between the nucleus and the cytoplasm, and at $\lambda = 2{,}650$ A the nucleus absorbs more strongly than the cytoplasm. The variation in absorption with changing wave length at a given point permits the identification of the local chemical composition. Such knowledge is an essential prerequisite for the understanding of basic processes of life.

QUESTIONS AND PROBLEMS

★1. Can you design a prism spectroscope through which light of a given frequency would pass without deviation?

2. In what way does a prism spectrograph resemble a camera?

3. Can there ever be overlapping between the spectra of the first and of the second order with a grating spectroscope?

4. Would there be any advantage in observing the grating spectrum of the third order instead of the first order?

5. How could you calibrate a prism spectroscope so as to be able to read off wave lengths of recorded lines?

6. How could you study infrared spectra beyond the region of photographically effective infrared?

7. Could you study stellar spectra without a slit?

8. What would an observer on the moon see in the sun's spectrum? How would the spectrum of moonlight observed on the surface of the earth differ from the sun spectrum observed from the moon?

9. When would you prefer a grating spectrograph to a prism spectrograph? When would you prefer the latter?

★10. A certain grating is such as to produce only one order of the visible spectrum. One notices, however, when viewing a monochromatic source, a series of very faint lines which look like several orders of spectrum produced by a grating of a larger grating constant. How could you explain this observation? Could it be a defect of the grating?

11. What do we hear when a buzzing bee passes our ear? How do you explain the sound effect?

12. Car A going east approaches car B which is going west. Each car moves with a speed of 30 m/sec with respect to the ground. Car A sounds a horn which emits a sound of frequency $f = 300$ vibrations per second. What frequency does the chauffeur of car B hear (*a*) before the cars meet? (*b*) after the cars pass each other?

13. An astronomer studies the sodium D line in the absorption spectrum of two adjacent stars. The wave length corresponding to this line as determined in the laboratory is 5,893 A. He finds, however, slightly different values for the wave length of the sodium line which appears in the stars' spectrum. For star A he finds $\lambda' = 5{,}800$, and for star B, $\lambda'' = 5{,}910$. What quantitative conclusion does he draw from this observation?

★14. Could you detect rotation of a distant *star* by means of the Doppler effect? (Remember that stars appear as points and not as disks in a telescope.)

15. What effect does the thermal motion of molecules have upon the width of spectral lines?

16. How can you determine the temperature of a star?

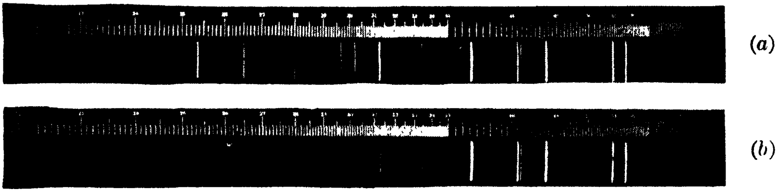
17. Can the Doppler effect be considered as evidence in favor of the wave theory of light?

★18. Suppose you project a real image of the planet Jupiter upon a vertical spectroscope slit so that the upper edge of the planet moves toward you, owing to the planet's rotation about its axis, and the lower edge away from you. How will the spectral lines appear on the photographic plate in relation to the orientation of the slit?

19. Draw a schematic diagram of a setup for spectroscopic examination of the content of a living cell making use of a reflecting microscope and a spectrograph.

20. What objection could be raised against observation of living cells with the reflecting microscope using ultraviolet light?

21. The photographs show the line spectrum of a mercury arc maintained (*a*) in a quartz tube, (*b*) in a glass tube. On the basis of comparison between these photographs draw a curve indicating qualitatively absorption of light in glass as compared with quartz as a function of wave length in the violet and ultraviolet region. (The scale divisions multiplied by 100 give λ in angstrom units.)



(*From Jenkins and White, "Fundamentals of Optics," McGraw-Hill.*)

22. A certain spiral nebula speeds away from our galaxy with a speed of 3×10^9 cm/sec. At what location of the wave-length scale of the spectroscope will the sodium *D* absorption line appear if its location in the solar spectrum is at $\lambda = 5{,}893$ A?

23. How fast must a source of yellow light ($\lambda = 5{,}800$A) move away from a mirror in order for the reflected light to appear red ($\lambda = 6{,}500$A) to an observer moving along with the source?

24. Which color of light is emitted with maximum intensity from the surface of a glowing body at 6000°C (about the temperature of the sun's surface)?

BIBLIOGRAPHY

Caspersson, T., Methods for the Determination of the Absorption Spectra of Cell Structures, *J. Roy. Microscop. Soc.*, Vol. 60, p. 8, 1940.

Barer, R., On the Use of the Reflecting Microscope in Microscopic Spectral Analysis, *Lancet*, Vol. I, p. 533, 1949.

CHAPTER 32

OSCILLATIONS OF ELECTRIC CHARGES

a. How does an a-c generator work?

b. What are the advantages of alternating as compared with direct current?

c. What is the meaning of the term "effective current"?

d. What is the relationship between effective and peak values of current and voltage? Does this relationship hold for alternating current of any "wave form"?

e. What is the justification for the vector representation of alternating currents and voltages? What is the advantage of this representation?

f. What is the phase relation between current and applied voltage (1) in a pure ohmic resistance? (2) in a pure inductance? (3) in a pure capacitance? (4) in an *RCL* circuit?

g. How do we define the impedance of a circuit?

h. What is the value of the impedance of (1) a pure resistance? (2) a pure inductance? (3) a pure capacitance? (4) an *RCL* series circuit?

i. How does the impedance depend on the a-c frequency for (1) a pure ohmic resistance? (2) a pure inductance? (3) a pure capacitance?

j. Does a condenser constitute a break (1) in an a-c circuit? (2) in a d-c circuit?

k. What is a displacement current?

l. Can there be an electric current in a region where no displacement of electric charges takes place?

m. Is the ratio of instantaneous voltage across a circuit element to instantaneous current necessarily equal to its impedance? For what type of circuit element is this ratio constant? For what types of circuit elements is it not constant?

n. What is meant by "resonance" in an a-c circuit?

o. What is resonance due to?

p. Can an *RL* circuit resonate?

q. Is the *LC* circuit the only circuit capable of resonance? Why or why not?

r. If you apply an alternating voltage of constant amplitude to an *RCL* series circuit, how do you expect the current to change as you vary the frequency from 0 to ∞?

In previous chapters we considered oscillations of neutral particles and saw how they can excite oscillations of neighboring particles, thus giving rise to the propagation of wave motion in solids, liquids, and gases. Certain phenomena observed in the case of light (interference effects) showed a typical pattern of behavior characteristic of waves and led us to the belief that light is due to vibratory motion of the particles of a hypothetical mechanically imperceptible medium, the "ether," which is capable of transmitting quasielastic waves as air transmits sound waves. In this and Chap. 33, we shall study oscillatory motion of electrically charged particles and shall try to determine whether or not vibrating electrical particles can give rise to a new type of wave motion different from sound and light. Our study will show that oscillations of electric charges do, indeed, generate so-called *electromagnetic waves.* These waves will be shown to be in many ways different from sound waves but not from light. In fact, we shall identify visible light and several other familiar radiations with electromagnetic waves.

This chapter is devoted to the study of alternating currents, *i.e.*, to the to-and-fro motion of electric charges in various electric circuits. Chapter 33 will pay attention primarily to the varying electric and magnetic fields which accompany the oscillations of electric charges. It will be seen in this chapter how the mathematical formalism and geometrical representation developed in connection with the mechanics of vibratory motion can be used for the description of alternating-current phenomena. This will demonstrate to us how a mathematical formalism developed in one field of physics may be applied to describe seemingly unrelated physical phenomena. We shall become acquainted with a new type of physical theories, mathematical theories, which are not based on a mechanical picture of a process and the essence of which is in a set of mathematical equations.

Alternating-current Generators

Let us begin with the consideration of oscillating electric charges. What would be the simplest way of producing a back-and-forth motion of electric charges? One might think of the following possibility: If we put sand in a test tube and shake it back and forth, the sand will move relative to the tube. Should not the same thing happen to the electrons when a piece of metal is shaken rapidly back and forth? Tolman has, indeed, shown that an alternating current is produced in a metallic conductor owing to the inertia of the electrons when the metal is made to oscillate rapidly. But this effect is very small, and the currents are difficult to detect. It is much easier to generate **alternating currents** (currents of varying magnitude and direction) by subjecting electrons or ions to an alternating electric field. One simple way of doing it would be to wave, say, a charged glass rod near a piece of metal or to make a conductor vibrate in a magnetic field, thus subjecting its conduction electrons to an alternating Lorentz field (see Chap. 24). The latter experiment is illustrated in Fig. 27.1. The microphone M is essentially a generator of alternating electric currents. This kind of a-c† generator is not a powerful source of electrical energy. The method of inducing an emf in a coil illustrated in Fig. 24.2 can, however, be used for the generation of large alternating currents. Figure 32.1 illustrates the principle of a practical a-c generator. C is a loop of wire which is made to rotate about the axis A in a magnetic field. The loop is not closed directly. Its ends are connected to the slip rings, which are tapped by the brushes B. These brushes press against the rings and are connected to the device to which the induced alternating current is to be supplied. In practice, one uses a coil of many turns instead of a single loop C, thus using all the emf's induced in the individual turns of the coil in series.

---

† The abbreviation of alternating current is a-c; d-c is an abbreviation of direct current, *i.e.*, a current of constant direction.

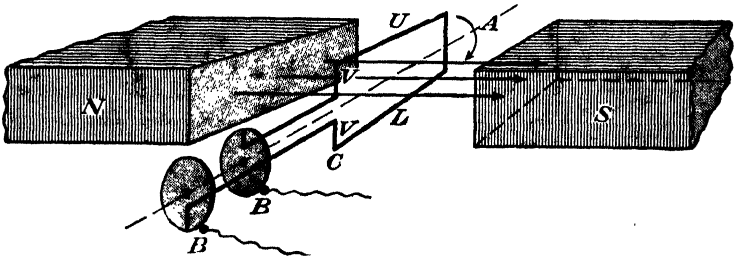


Fig. 32.1. Scheme of an a-c generator.

How does the current induced in this generator vary as a function of time? This can be inferred with the aid of Fig. 32.2a. Only the sections U and L of the loop wires (Fig. 32.1) move through the magnetic field so as to cut its lines of force. Thus no emf is induced in the sections V, all of the induced emf being due to the motion of U and L. The motion of these parts of the loop is a circular one and is shown in side view in Fig. 32.2a. The instantaneous value of the emf induced

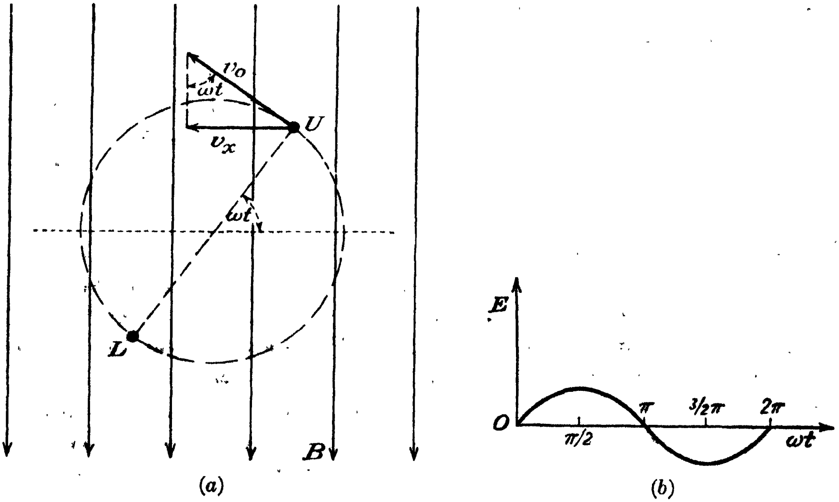


Fig. 32.2.

in a straight wire (*e.g.*, wire U or L) is given by the expression†

$$E' = BLv_x \times 10^{-8} \text{ volt} \tag{24.6}$$

From Fig. 32.2*a* the component of v_0 perpendicular to the magnetic field is seen to be

$$v_x = v_0 \sin \omega t$$

Hence the alternating emf induced in the wire U or L is given by

$$E' = BLv_0 \sin \omega t \times 10^{-8} \text{ volt}$$

This expression can be simplified by considering that v_x reaches its maximum value $v_x = v_0$ when the line UL reaches the vertical position. Then $\omega t = 90°$ and $\sin \omega t = 1$. At that moment the instantaneous value of the induced emf is larger than at any other time during the cycle. If we introduce $E_0 = BLv_0 \times 10^{-8}$, we can simplify our expression for E' by writing

$$\boxed{E' = E_0 \sin \omega t} \tag{32.1}$$

We see that the induced emf is a sine function of time. Figure 32.2*b* shows E' as a function of the angle ωt, where ω is the angular speed of the rotating coil. If we connect the generator to an ohmic resistor (a conductor of negligible self inductance), we find that the current is proportional to E' and hence is also a sine function of time. Since the sine of the angle ωt assumes periodically negative (for $2\pi > \omega t > \pi$) as well as positive (for $\pi > \omega t > 0$) values, the emf and the current must periodically change direction in the conductor.

Advantages of Alternating Currents

It is obvious that computations relating to alternating emf's and currents must be much more complicated than the simple application of Ohm's law to constant-current circuits. Why then are alternating currents used in preference to direct currents in spite of their greater complexity? The primary reason for it is their "transformability." One of the great advantages of electrical power is due to the fact that it can be transported simply through wires over long distances, allowing one to tap naturally available power, say of a waterfall, to feed machinery located hundreds of miles away. This transport of electrical power necessarily involves loss of electrical energy due to generation of heat in the long wires. If R is the total resistance of the wires which conduct i from the power plant to a city, the heat loss in the wires is given by i^2R. The total power supplied by the power plant is $P = V_g i$, where V_g is the potential difference across the generator terminals (the power dissipation inside the generator is not considered here). We see that the same power could be delivered at a variety of values of V and i while the product $P = Vi$ remains constant. We could, for instance, use a low-voltage generator and a heavy current or low current at a high voltage. The advantage of the latter procedure for the transport of electric power is obvious. If we reduce the current i by a factor of 10, the i^2R loss in the transmission line will be cut down by a factor of 100. *Thus the use of very small currents allows us to use relatively thin and* hence cheap lead wires. But the transmission of power at low current requires a power source of high voltage which is difficult to construct for d-c generation but is relatively simple if an a-c voltage is used. We shall see below how alternating voltage may be transformed up or down. Alternating voltages of the order of 300,000 volts can be obtained from the secondary winding of large **power transformers** whose primary coil is connected to the relatively low voltage output of an a-c generator. The power is supplied at high voltage and low current from the secondary coil to a substation near the consumer, where it is transformed down to a safe low voltage, say 110 volts, and distributed to local power outlets. The same method of obtaining alternating voltages of different magnitude is also used in various appliances. For instance, radio sets contain transformers which reduce the line voltage to values of 2.5 or 6.3 volts to heat the filaments of radio tubes and others which transform it up to voltages of a few hundred volts to supply the "plate voltage" for various tubes.

---

† $B = \mu H$, magnetic induction; L = length of wire; v_x = horizontal velocity component.

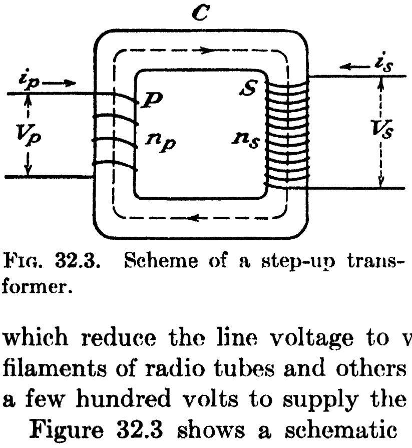


FIG. 32.3. Scheme of a step-up transformer.

Figure 32.3 shows a schematic diagram of a **transformer.** C is an iron core upon which two coils are wound: P (primary coil) and S (secondary coil). n_p and n_s represent the number of turns of these coils. The dashed line inside the core represents one of the magnetic flux lines which are confined mostly to the interior of the core. The magnetic flux is generated by the current i_p through the primary coil. This alternating magnetic flux induces the same emf in each turn of the primary as well as of the secondary coil. Hence, the total voltage across the terminals of each coil will be n times as large as the voltage per turn, where n is the number of turns of the coil. The ratio of the primary voltage V_p to the secondary voltage V_s will therefore be equal to the turn ratio of the coils:

$$\frac{V_p}{V_s} = \frac{n_p}{n_s}$$

In the illustrated case, the secondary coil has more turns, so that the transformer acts as a "step-up transformer." Such transformers are used, for instance, as a high-voltage source for X-ray machines. "Step-down transformers" in which $n_s < n_p$ are also widely used (for instance, filament transformers for radio tubes).

Effective Values

Let us now consider the power supplied by an alternating current to a circuit element. In the case of constant current the power is given by the expression $P = Vi$ where $i = Q/t$. In the case of variable currents the *instantaneous* current is defined by the quotient $i' = \frac{\Delta Q}{\Delta t}$ where ΔQ is the charge moving through a cross section of the circuit in the very short time interval Δt. We shall define $P' = V'i'$ as the **instantaneous power,** where V' and i' are the instantaneous values of the current and of the voltage across the conductor. The meaning of the instantaneous power is then as follows: $P' = \frac{\Delta W}{\Delta t}$, where ΔW is the amount of energy delivered during the very short time interval Δt. Only if Δt is short enough can the value of P' be considered the instantaneous value.

We shall use the following symbols in our discussion of alternating currents:

1. Instantaneous values of current and voltage: i', V'.
2. Maximum values of current and voltage: i_0, V_0.
3. Effective values of current and voltage (see page 615): i, V.

Let us determine the average power delivered by an alternating current to a resistor R. Figure 32.4*a* shows the current flowing through R. We can obtain the instantaneous power $P' = i'^2R$ as a function of time by squaring the ordinate of graph 32.4*a* and plotting $(i')^2R$ against time in Fig. 32.4*b*. From this graph it is easy to find the **average power,** *i.e., the energy delivered during the period of a cycle divided by the duration of the cycle.* The energy delivered to our resistor during a short time interval Δt is given by $P'\,\Delta t = (i')^2R\,\Delta t$. This expression has a simple geometrical meaning as can be seen from Fig. 32.4*b*: The area of the narrow dotted rectangle B can be considered as one of a multitude of such rectangles which can be drawn side by side to approximate the actual $R(i')^2$ curve by a "staircase" (Fig. 32.5). If the values of Δt are chosen sufficiently small, the area under the staircase will be practically equal to the area located under the P' curve. The area under the staircase is the sum of the areas $\Delta t(i')^2R$. Each of these small areas has the meaning of the total energy supplied during Δt seconds. Hence, the area under the staircase has the meaning of the total energy supplied over the entire cycle if the "staircase" extends over the full length of the period, *i.e.*, from A to E in diagram 32.4*b*. This energy (*i.e.*, this area) divided by the duration of the cycle is, by definition, numerically equal to the average power. The area under the curve can be found simply by the following artifice. We draw a horizontal line (dashed line H of Fig. 32.4*b*) so that the area of the hilltops which it cuts off (shaded areas) is equal to the area of the dales below it (crosshatched areas under H). Then the area of the rectangle $AECD$ is equal to the original area under the $(i')^2R$ curve. This area is $A = YT$. Because of the symmetry of the curve, it is clear that $Y = \frac{1}{2}(i_0^2R)$. Thus we obtain for the average power

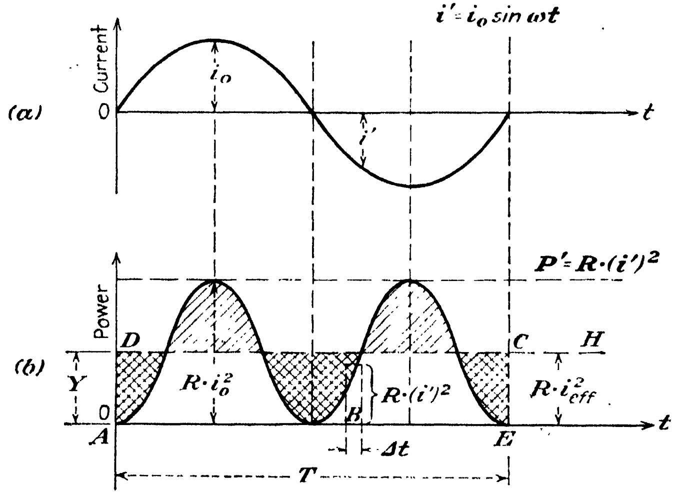


FIG. 32.4*a*. A sinusoidal alternating current. *b*. Power dissipation in a resistor traversed by alternating current.

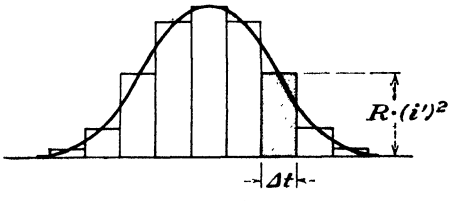


FIG. 32.5.

$$P = \frac{\text{energy input in a cycle}}{\text{duration of cycle}} = \frac{YT}{T} = Y = \frac{1}{2}i_0^2R$$

The same power could have been delivered to our resistor R by a constant current i. This power would have been Ri^2. Since this power is to be equal to the power delivered by the alternating current, we can write

$$Ri^2 = \tfrac{1}{2}i_0^2R$$

From this follows

$$\boxed{i = \frac{i_0}{\sqrt{2}}} \tag{32.2}$$

It is convenient to introduce a new concept, namely, that of an **effective current.** *We define as the effective value of an alternating current the value of a "direct" (i.e., constant) current which would supply the same power to a resistor as the alternating current does on the average.* Equation (32.2) makes it possible to compute the effective current value from the peak value, i_0. Similarly, we can derive the following relation between effective and peak voltage:

$$\boxed{V = \frac{V_0}{\sqrt{2}}} \tag{32.3}$$

The effective value of an alternating voltage is that value of a constant voltage which would supply the same power to a resistor as the alternating voltage does on the average.

We can use the same power formulas with alternating current as we did with direct current ($P = i^2R = V^2/R = iV$) provided i and V are taken to mean effective values.

Vector Representation of Alternating Currents

We have considered so far an alternating current created by a single generator. It is conceivable that two or more generators could be put in series like batteries. Then a possible complication arises which does not occur with batteries: The generated emf's and hence the currents may not be in phase. Figure 32.6 illustrates what is meant by this statement. This figure shows a snapshot of three generators (which may be connected in series). The coils of A and B happen to be oriented alike, so that the emf's induced in them will reach a maximum value at the same time. These emf's are said to be *in phase.* The generator C has just passed the orientation in which the emf induced in it is a maxi-

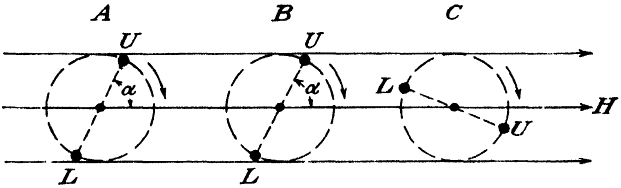


FIG. 32.6.

mum. It has passed this position before coils A and B reach it. The phase of C is said to be *leading the phase* of A and B. The emf induced in A and B is *lagging* the emf in C by an angle of about 90° or $\pi/2$ radians.

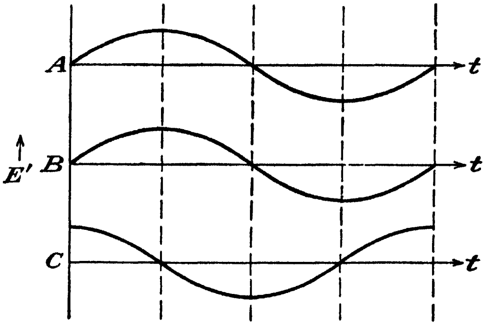


FIG. 32.7.

The ordinates of Fig. 32.7 show how the emf induced in A, B, and C varies with time.

What would the current look like in a circuit energized by two generators which are not in phase such as B and C? In order to facilitate the answer to this question, we shall introduce a geometrical representation of alternating currents similar to the one used in connection with harmonic motion in Chap. 25.

Let us consider equation (32.1) for the instantaneous emf of a generator. Let us assume that the generator is connected to a conductor, the instantaneous current through which can be found by Ohm's law:

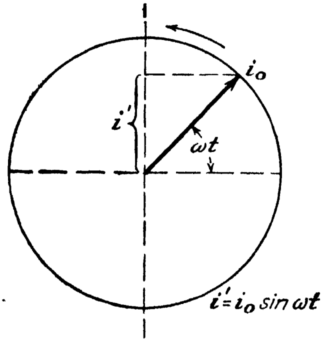


FIG. 32.8.

$$\boxed{i' = \frac{E'}{R} = \frac{E_0}{R} \sin \omega t = i_0 \sin \omega t} \qquad (32.4)$$

The expression $i' = i_0 \sin \omega t$ is analogous to $x = x_0 \sin \omega t$ for the displacement of a harmonic oscillator. In Chap. 25, we saw that x could be considered as the length of the shadow (projection) of a rotating vector of constant length x_0 (Fig. 25.2). We can similarly represent i' geometrically as the projection (upon the *vertical* axis) of a vector of constant length i_0, rotating with the angular speed ω (Fig. 32.8).

Let us now suppose that two alternating currents (of equal frequency) which are out of phase flow through our circuit. The two individual currents can then be represented by the projections i'_1 and i'_2 of two rotating vectors $(i_1)_0$ and $(i_2)_0$ which enclose an angle θ. This angle represents their *phase difference* (Fig. 32.9).

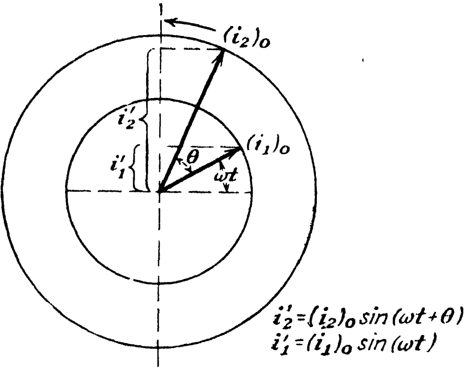

$(i_2)_o$
i'_2
i'_1
θ
$(i_1)_o$
ωt
$i'_2=(i_2)_o \sin(\omega t+\theta)$
$i'_1=(i_1)_o \sin(\omega t)$

FIG. 32.9.

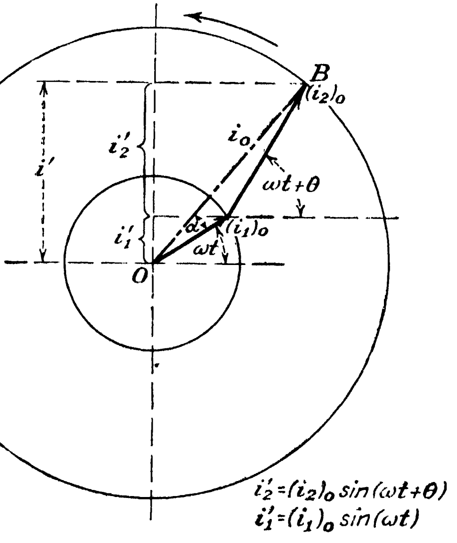


FIG. 32.10.

The instantaneous value of the resultant current is given by the sum of the instantaneous component currents:

$$i' = i'_1 + i'_2$$

i' is the sum of two projections, and we can consider it to be the projection of another rotating current vector i_0. How can we find i_0 if $(i_1)_0$ and $(i_2)_0$ and their phase difference are given? Consider Fig. 32.10. It has been redrawn from Fig. 32.9. The $(i_1)_0$ vector is shown in the same position as in Fig. 32.9. The $(i_2)_0$ vector has been added to it vectorially. The projections of these two vectors are marked as $(i_1)'$, and $(i_2)'$ in the diagram. We see that the sum of these projections is equal to the projection of the OB vector (dot-dashed line). Hence, $\overrightarrow{OB}$ can be considered as the sought i_0 vector whose projection gives at any moment the instantaneous value i' of the resultant current. We see from the figure that *the phase difference* α *between the resultant* i' *and the component* $(i_1)'$ *is different from the phase difference* θ *between the component currents* $(i_1)'$ *and* $(i_2)'$. The vectorial method of representation of alternating currents and voltages will prove very helpful in subsequent derivations of important relationships.

We shall now turn to the consideration of the phase relationship between current and voltage in various types of circuits.

The Pure Ohmic Circuit

In general, when we apply an alternating potential difference across a conductor, observations show that the instantaneous current is not exactly in phase with the applied voltage but rather lags behind it more or less, and the effective value of the current is not given by $i = V/R$ but rather by $i = V/Z$ where $Z > R$. However, *there is a large class of conductors for which, for all practical purposes,* $Z = R$ *and* i' *is in phase with* V'. **Such conductors we shall call pure "ohmic resistance"** (*definition of ohmic resistance*).

The Purely Inductive Circuit

The departure of a conductor from the behavior of a pure ohmic resistance is in most practical cases due to its self-inductance. In order to understand the general case we shall consider first a special ideal case of a superconductor possessing zero ohmic resistance and a self-inductance L. It was shown in Chap. 24, following the discussion of equation (24.5), that, although it is impossible to establish a constant potential difference across a superconducting coil by sending a constant current through it, an alternating potential difference can be maintained between the terminals of the coil when an alternating current is sent through it. The variation of the magnetic field, which is associated with the current, induces an emf in the coil which, according to Lenz's rule, opposes the *changes* in current that lead to the induction of this emf. It is because of this feature that this emf is frequently referred to as a *counter emf.* Its magnitude is given by equation (24.5): $e' = -L\frac{\Delta i'}{\Delta t}$. Figure 32.11 illustrates how the phase relationship between the coil current and the emf induced in the coil can be established graphically. Figure 32.11*A* shows the graph of the current $i' = f(t)$. To find the induced emf as a function of time we find first $\left(\frac{\Delta i'}{\Delta t}\right)$ as a function of time.

The meaning of this expression is illustrated by Fig. 32.12. The slope of the line T is given by $\tan \theta = \frac{\Delta i'}{\Delta t}$ from the triangle ABC. As Δt is chosen smaller and smaller, the point B approaches point A more and more and finally the line T practically coincides with the tangent to the curve at point A. Thus, the *quotient* $\frac{\Delta i'}{\Delta t}$ *can be considered as the slope of the tangent* (and hence as the slope of the curve) at point A (*i.e.*, at the time t) provided Δt is chosen infinitesimally small. Thus we could plot the $\frac{\Delta i'}{\Delta t}$ curve

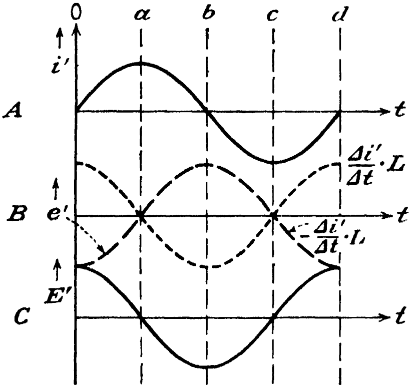

0
a
b
c
d
i'
A
t
$\frac{\Delta i'}{\Delta t}\cdot L$
B
e'
t
$-\frac{\Delta i'}{\Delta t}\cdot L$
E'
C
t

FIG. 32.11.

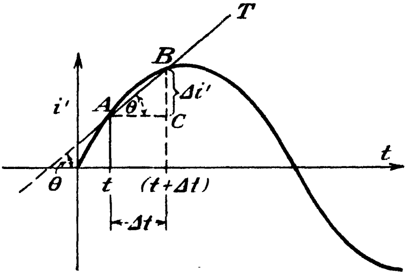


FIG. 32.12.

by drawing the tangent to the curve of Fig. 32.11A at different values of t, measuring its slope, and plotting the value of the slope as ordinate in Fig. 32.11B at the appropriate value of t (dotted curve). Actually an additional operation, namely, multiplication by L, has been performed at the same time, yielding the *dotted curve*, representing $L\frac{\Delta i'}{\Delta t}$ in Fig. 32.11B. It is easy to ascertain the position of a few points of curve B as follows: The tangent at the maximum and minimum of curve A is horizontal; hence the slope at the abscissae a and c is $\tan \theta = 0$, and the dotted curve should cross the t axis at a and c. On the other hand, the slope at abscissae O and d is positive and steeper than at any other point of the curve. There should be maxima in the dotted slope diagram (Fig. 32.11B) at these points. Since the slope of the curve A at abscissa b is negative and algebraically smaller than at any other point, a minimum of slope should lie at b of the dotted curve in Fig. 32.11B.

To obtain the desired value of the self-induced emf we have merely to multiply $\frac{\Delta i'}{\Delta t}L$ by (-1), which means we have to invert the dotted curve. This leads to the dashed e' curve in Fig. 32.11B.

In most practical cases the alternating current flowing through the coil is the result of an alternating emf applied to it by connecting the coil to some kind of generator. *We have to distinguish between the applied emf E' and the self-induced emf e'.* Let us establish now the phase relation between the coil current and *applied* emf. Let us assume now that our coil is not a superconductor but rather that it possesses a resistance R. For clarity, this resistance has been drawn in Fig. 32.13 connected in series to coil L, which possesses only self-inductance. G is an a-c generator whose emf will be designated as the applied emf E'. The alternating current, which is maintained by G in the circuit, induces an emf in the coil L, which can be considered as a second source of emf connected in series with G. The instantaneous current through R will then be given by Ohm's law as follows:

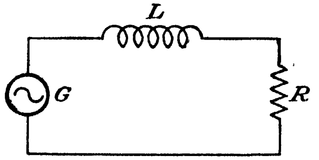


Fig. 32.13. An a-c circuit consisting of generator, resistor, and inductance.

$$\text{Total emf} = E' + e' = E' + \left(-L\frac{\Delta i'}{\Delta t}\right) = i'R \qquad (32.5)$$

If the coil happens to have a negligible resistance, the product $i'R$ is zero, and we get

$$\boxed{E' = -e'} \qquad (32.6)$$

This means that **the emf of self-induction in the superconducting coil L (counter emf) is equal and opposite to the applied emf,** *i.e.*, the two emf's are in phase opposition.

Now we can very simply draw the graph of E' as a function of time. We have merely to invert the dashed e' curve of Fig. 32.11B, which makes it identical with the dotted curve, and yields curve C.

Examination of curves 32.11A and C reveals an important fact. **The applied voltage and the current are 90° out of phase in a pure inductance, the current lagging behind the applied voltage**; *i.e.*, the voltage reaches its maximum at a smaller value of t than the current.

Inductive Impedance

We shall now derive a relation between the applied emf E' and the observed current $i' = i_0 \sin \omega t$ to which it gives rise. We know that $E' = -e'$ and that $e' = -L\frac{\Delta i'}{\Delta t}$ (for $R = 0$).

Hence,
$$E' = L\frac{\Delta i'}{\Delta t} \qquad (a)$$

In Fig. 32.11 we have obtained E' graphically on the basis of this equation. We shall now attempt to find it analytically. We can do this without computation by making use of the formal analogy between the equations of harmonic motion and those pertaining to alternating currents. Both can be interpreted in analogous fashion by means of the rotating vector diagram. The equations of harmonic motion have been solved previously, and we can simply transfer the results of those solutions to the case of alternating current by a proper interchange of symbols. From a comparison of Fig. 25.2 with Fig. 32.8 it follows that we obtain the electrical vector diagram from the mechanical diagram by replacing the rotating x_0 vector by the rotating i_0 vector. From a comparison of the expressions (B) and (b) for the projections of the mechanical and electrical vectors (see below), we infer that the electrical equation is obtained from the mechanical equation by replacing x by i' and x_0 by i_0. Equation (C) is a mathematical consequence of equation (B). By formal analogy equation (c) must be a mathematical consequence of equation (b).

Harmonic motion	Alternating currents
(B) $x = x_0 \sin \omega t$	(b) $i' = i_0 \sin \omega t$
(C) $v = \dfrac{\Delta x}{\Delta t} = \omega x_0 \cos \omega t$	(c) $\dfrac{\Delta i'}{\Delta t} = \omega i_0 \cos \omega t$

Equation (c) is the desired expression for $\frac{\Delta i'}{\Delta t}$ which we want to substitute in above equation (a). Substituting (c) in (a), we get

$$\boxed{E' = L\, \omega i_0 \cos \omega t} \qquad (32.7)$$

E' is seen to vary periodically. It reaches the value zero when $\cos \omega t = 0$ and for $\cos \omega t = 1$ a maximum value of

$$\boxed{E_0 = \omega L i_0} \qquad (32.8)$$

If we write the equation (32.8) as follows:

$$\boxed{\frac{E_0}{i_0} = \omega L} \qquad (32.8a)$$

we obtain an expression which is analogous to the definition of a resistance: $E/i = R$. We call this ratio of maximum voltage over maximum current the

$$\boxed{\text{Inductive impedance: } Z = \omega L}^{\dagger}$$

---

† The conventional term for ωL is "inductive reactance." The word reactance is not used in this book to avoid overloading the student with too many new terms.

Since the effective values of current and voltage are proportional to the maximum values [see equations (32.2) and (32.3)] we can write just as well

$$\boxed{\frac{E}{i} = \omega L} \tag{32.8b}$$

It is important to realize that E and i do not stand for instantaneous values of emf and current. A brief glance at Fig. 32.11 will teach that the ratio E'/i' is not constant. For instance, at $t = 0$, E' is finite (from curve C) while $i' = 0$. Hence $E'/i' = \infty$. On the other hand, a quarter of a cycle later (at the abscissa a), $E' = 0$ and i' is finite, so that $E'/i' = 0$.

Equations (32.8a) and (32.8b) have the form of Ohm's law. They state that we can obtain the maximum voltage drop across a coil whose coefficient of self-inductance is L by multiplying the maximum current by the inductive impedance ωL and that a similar relation exists between the effective values of voltage and current and the impedance. **Equations (32.8a) and (32.8b) state that the inductive impedance ωL limits the flow of alternating current through a self-inductance just as the resistance R limits the flow of direct current through an ohmic resistance.**

Since both R and ωL **are defined as a ratio between voltage and current, they are measured in the same units, namely, in ohms.**

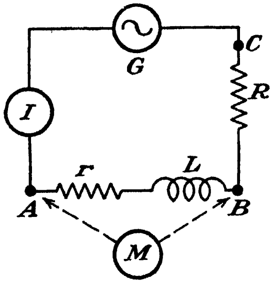


Fig. 32.14.

The following experiment illustrates the physical meaning of the preceding equations (Fig. 32.14). G is an a-c generator of variable frequency; r is the resistance of the coil L as determined by d-c measurements. R is a resistor, and M is an a-c voltmeter which measures d-c potential drops as well as effective values of alternating voltage. I is an a-c ammeter, which measures direct currents as well as the effective value of alternating current.

Suppose $r = 2$ ohms, $L = 5$ henrys, $R = 100$ ohms. Then if we replace G by a d-c battery and send a *constant* (direct) current of $i = 2$ amp through the circuit, we measure the following voltages.

Between points A and B: $V_{AB} = iR_{AB} = ir = 2 \times 2 = 4$ volts
Between points B and C: $V_{BC} = iR = 2 \times 100 = 200$ volts

If we now pass an *alternating* current of the same effective value of

$i = 2$ amp and of $\omega = 2\pi f = 100$ rad/sec, our readings of the effective voltage will be

$$V_{BC} = iR = 2 \times 100 = 200 \text{ volts}$$

as before and (*approximately*) $V_{AB} \simeq i(\omega L) = 2(100 \times 5) = 1{,}000$ volts. In computing the last value we neglected the resistance r against the much larger impedance ωL.

We see that, when alternating current is flowing in our circuit, the effective voltage distribution is quite different from the case of d-c flow. Whereas in the case of direct current only an insignificant potential difference of $V_{AB} = 4$ volts appeared across the coil L and a potential difference of $V_{BC} = 200$ volts across R, in the case a-c flow the effective voltage $V_{AB} > V_{BC}$. Moreover, it is in our power to make V_{BC} insignificantly small as compared with V_{AB}. The latter value depends on ωL and by choosing a higher frequency we can make ωL as large as we please.

Capacitive Impedance

Heretofore, when speaking of an electric circuit, we were making the tacit assumption that it had to be a closed chain of electric conductors.

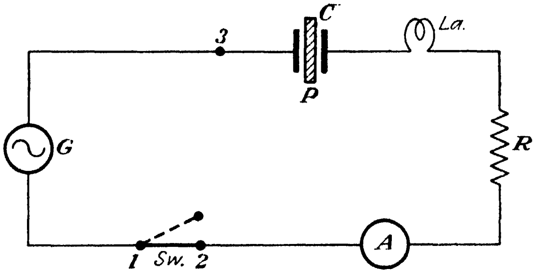


FIG. 32.15. An a-c circuit consisting of generator, capacitance, and resistance.

We know that a break in the continuity of this chain would interrupt the flow of direct current, and we are tempted to assume that the same would happen to the flow of alternating current. The following experiment shows, however, this assumption to be false (Fig. 32.15). G is a generator of an alternating emf. A is an a-c ammeter, R is a resistor, and La is a lamp which lights up when a high enough current flows through it. C is a condenser consisting of two parallel plates and a glass plate P between them. The condenser constitutes a dielectric gap in the chain of electric conductors, but nevertheless, if the capacity of C is high enough, the lamp is lit and the ammeter A registers the flow of alternating current. If, however, G is replaced by a d-c generator, the ammeter reads zero current and the lamp remains dark.

The condenser C appears to conduct an alternating current. The glass plate P improves this conduction, but even when P is pulled out, the current persists although it is somewhat weakened. A vacuum between the condenser plates does not interrupt the flow of alternating current through this circuit, and yet observations show that no electrical charges are transported through the space between the plates! How, then, can we speak of the flow of an electric current in our circuit if there is no flow of charges through one of its sections?

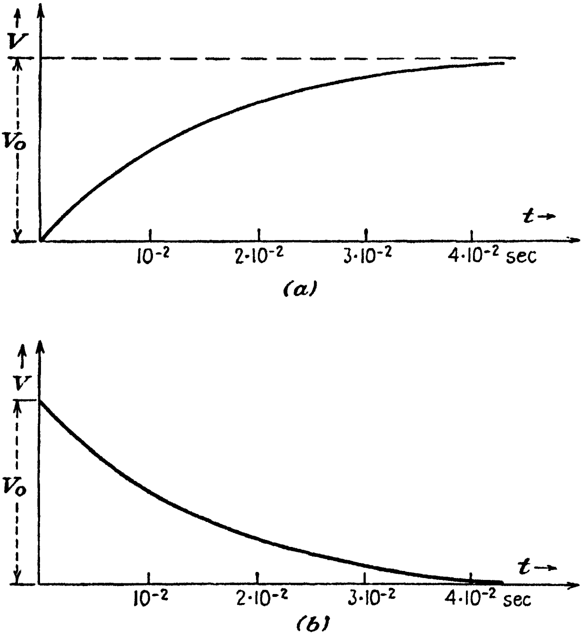


FIG. 32.16*a*. Voltage across a capacitor which is being charged through a resistor. *b*. Voltage across a capacitor which is being discharged through a resistor.

The following observation leads to a clarification. If we replace G by a d-c battery, we notice that the lamp flashes momentarily as we close the switch *Sw*. The lamp can be made to flash again by connecting point 2 to point 3 by a copper wire after the switch *Sw* has been opened. A cathode-ray oscillograph operating as a quick-acting voltmeter connected to the two condenser plates helps explain this observation. Figure 32.16*a* shows the oscillogram of the potential difference across the condenser plates as it changes after the switch *Sw* has been closed. We see that it rises rapidly at first, approaching more and more slowly a limiting value, which is equal to the potential difference of the d-c battery. As the condenser is being charged, electric charges accumulate on its plates. Connecting points 2 and 3 discharges the condenser through the lamp, resistor, and ammeter. Figure 32.16*b* shows the dropping voltage as a function of time during the period of discharge.

During the charging process a current flows through the wires which connect the battery to the condenser, until the potential difference between the plates becomes equal to the battery voltage. This temporary charging current is evidently responsible for the flashing of the lamp. When points 2 and 3 are bridged conductively, the electrons flow from the negative condenser plate to the positive plate, passing through the lamp and making it flash again. We see that, if we could close the switch and then open it, at the same time rapidly connecting 2 to 3, we could make the lamp glow practically continuously by repeating this operation many times in rapid succession. It would be more efficient simply to keep reversing the polarity of the battery instead of proceeding as suggested above. Then we would first move electrons, say, to the left plate and at the same time extract them out of the right condenser plate, thus creating a momentary charging current. When we reverse the connections to the poles of the battery, the electrons begin to flow toward the right plate, which was previously robbed of some electrons, while the plate which had an electron surplus sends electrons back to the positive pole of the battery until it has become positively charged, *i.e.*, has an electron deficit. Thus we can maintain an alternating (though nonsinusoidal) charging current in the wires leading to the condenser and keep the lamp lit. An essentially similar process is to be expected in case an alternating sinusoidal voltage is applied to the condenser. Thus we can consider a capacitor as a circuit element which "conducts" alternating currents. Does a capacitor impede the flow of an alternating current and determine its value for a given alternating voltage applied to its plates as a self-inductance limits the flow of alternating current? We shall see that it does and shall derive an expression for the capacitive impedance through a consideration similar to that which yielded the expression for the inductive impedance.

The voltage V across a capacitor is related to the charge stored in it by the following equation: $Q = CV$, where the proportionality constant C is called the capacitance of the condenser (see Chap. 15). We see thus that the curves of Fig. 32.16*a* and *b*, which show how the voltage (of a condenser which is being charged or discharged) changes, also indicate how the charge Q on the condenser varies during this process. The instantaneous charging (or discharging) current can be defined as $i = \frac{\Delta Q}{\Delta t}$, that is, the tiny increment of charge divided by the short time interval during which it leaks into or out of the condenser. Because of the above relation between Q and V we can write

$$\boxed{i' = \frac{\Delta Q}{\Delta t} = \frac{\Delta(CV)}{\Delta t} = C\frac{\Delta V}{\Delta t}} \qquad (32.9)$$

Thus we see that a change in voltage across a condenser indicates the flow of a charging current which is proportional to the rate of change $\frac{\Delta V}{\Delta t}$. In Fig. 32.16, $\frac{\Delta V}{\Delta t}$ has the geometrical meaning of the slope of the curve. We see that the initial current according to both curves is high and that it gradually falls off to zero. i becomes zero when V becomes constant $\left(\frac{\Delta V}{\Delta t} = 0\right)$.

Let us now suppose that we apply a sinusoidal alternating voltage $V' = V_0 \sin \omega t$ to our condenser of capacitance C. What will be the charging current? The instantaneous value of the current is given by the equation (32.9), if V' stands for the instantaneous voltage across the condenser: $i' = C\frac{\Delta V'}{\Delta t}$. We simplify the solution again by making use of the formal analogy between our electrical equations and the equations of harmonic motion

Mechanical equations	Electrical equations
$(A)\ x = x_0 \sin \omega t$	$(a)\ V' = V_0 \sin \omega t$
$(B)\ \frac{\Delta x}{\Delta t} = (x_0\omega) \cos \omega t$	$(b)\ \frac{\Delta V'}{\Delta t} = (V_0\omega) \cos \omega t$

Again, as in the case of the inductance, equation (b) is a mathematical consequence of equation (a), from which it follows in the same fashion as equation (B) follows from equation (A).

Substituting (b) in (32.9) we obtain

$$\boxed{i' = C\frac{\Delta V'}{\Delta t} = CV_0\omega \cos \omega t} \qquad (32.10)$$

i' is seen to reach a maximum value of i_0 for $\cos \omega t = 1$. For this special case equation (32.10) becomes

$$\boxed{i_0 = V_0(\omega C)} \qquad (32.11)$$

or

$$\boxed{\frac{V_0}{i_0} = \frac{1}{\omega C}} \qquad (32.11a)$$

or, using the effective values,

$$\boxed{\frac{V}{i} = \frac{1}{\omega C}} \qquad (32.11b)$$

The analogy between equations (32.11) and (32.8) is apparent. The quantity $1/\omega C$, which we designate as the **capacitive impedance,** has the same significance in the case of a capacitor as ωL in the case of an inductance coil. **The capacitive impedance limits the flow of alternating current through a condenser just as the resistance limits the flow of current in a d-c circuit.**

Since $1/\omega C$ is determined similarly to R and ωL as a ratio of voltage over current, it is measured in the same units, namely, in ohms.

Since $\omega = 2\pi f$ is in the denominator of the capacitive impedance, we see that it decreases as the frequency goes up. *Thus a condenser transmits alternating currents better and better as the frequency is increased, whereas an inductance transmits alternating current less and less. The conductive behavior of a pure ohmic resistance, on the other hand, is independent of frequency.*

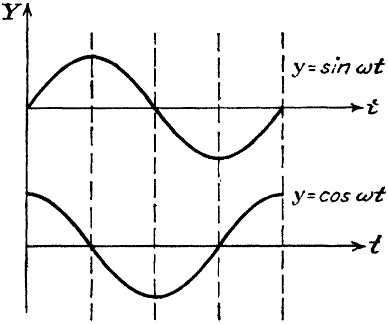


FIG. 32.17.

Figure 32.17 shows a plot of the functions sin ωt and cos ωt. We see that the phase of the cosine function is ahead of the sine function by 90°. Hence, equations (a) and (32.10) imply that **the current flowing through a capacitor leads the capacitor voltage V' by 90°.**

The *RCL* Circuit

Finally, we shall consider a series connection of three types of circuit elements: a resistor R, an inductance L, and a capacitor C (Fig. 32.18).

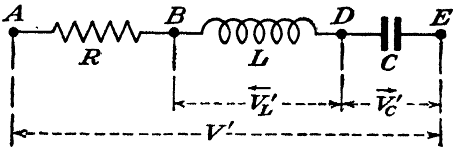


FIG. 32.18. An *RCL* circuit.

The current is in phase in all of these elements but the voltages between their terminals are out of phase.

$V'_R = V'_{AB}$ is in phase with the current.
$V'_L = V'_{BD}$ leads the current by 90°.
$V'_C = V'_{DE}$ lags the current by 90°.

The maximum potential difference between A and E is not equal to the algebraic sum of the maximum voltages across the individual circuit elements. The *instantaneous* voltage, however, is given by the sum of the instantaneous voltages across R, L, and C:

$$V' = V'_R + V'_L + V'_C \tag{1}$$

The instantaneous values of voltage are given by expressions derived in this chapter:

$$V'_R = Ri' = Ri_0 \sin \omega t \quad [i' \text{ substituted from (}b\text{) preceding (32.7)}] \tag{2}$$
$$V'_L = (V_L)_0 \cos \omega t = \omega L i_0 \cos \omega t \quad [\text{compare with (32.7) and (32.8)}] \tag{3}$$
$$V'_C = -(V_C)_0 \cos \omega t = -\frac{1}{\omega C} i_0 \cos \omega t \quad [(V_C)_0 \text{ follows from (32.11)}] \tag{4}$$

The negative cosine function has been used, since we know that V'_C and V'_L are in phase opposition.

Substituting equations (2), (3), and (4) in (1) we obtain

$$V' = Ri_0 \sin \omega t + \left(\omega L - \frac{1}{\omega C}\right) i_0 \cos \omega t \tag{5}$$

To simplify this equation, we set

$$\left(\omega L - \frac{1}{\omega C}\right) = X \tag{6}$$

Thus we can write

$$V' = Ri_0 \sin \omega t + Xi_0 \cos \omega t \tag{5a}$$

The sum expressed by equation (5*a*) can be represented geometrically as follows: Imagine two rods (vectors) of length (Ri_0) and (Xi_0) attached to each other rigidly at right angles as shown in Fig. 32.19 and revolving about point O as shown. The projections of these vectors upon the Y axis are, as can be read off from the diagram, $(Ri_0) \sin \omega t = V'_R$ and $(Xi_0) \cos \omega t = V'_X$. Their sum is designated as V'. V' is seen to be the projection of the hypotenuse $OP = V_0$ of the triangle whose sides are the vectors Ri_0 and Xi_0. V' is thus the variable projection of this V_0 vector which has the meaning of the maximum value which V' can achieve during a cycle; in other words, V_0 is the peak value of the potential difference V' between the points A and E in Fig. 32.18.

From the above-mentioned right triangle in Fig. 32.19 we obtain

$$V_0^2 = (Ri_0)^2 + (Xi_0)^2 \tag{7}$$

Let us now introduce the concept of the **impedance** Z of the circuit branch AE. We shall define it as the factor by which we have to multiply the maximum current i_0 in order to obtain the maximum voltage $(V_{AE})_0 = V_0$:

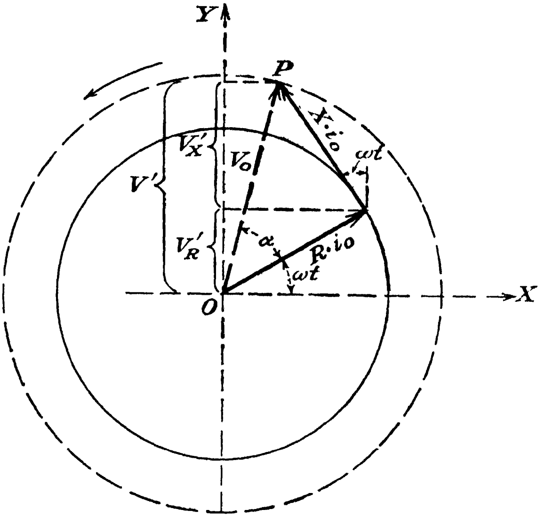


FIG. 32.19.

$$\boxed{V_0 = Zi_0} \tag{32.12}$$

Substituting equation (32.12) in equation (7) we obtain

$$(Zi_0)^2 = (Ri_0)^2 + (Xi_0)^2$$

Dividing this equation by i_0^2, we get

$$Z^2 = R^2 + X^2$$

or

$$Z = \sqrt{R^2 + X^2}$$

Substituting for X its original value from equation (6), we finally obtain

$$\boxed{Z = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \tag{32.13}$$

This suggests that the impedance can be represented graphically as the resultant of two vectors $\vec{R}$ and $\overrightarrow{[\omega L - (1/\omega C)]}$, as shown in Fig. 32.20. If $\omega L < 1/\omega C$, the vertical $[\omega L - (1/\omega C)]$ arrow points downward.

The angle α in Fig. 32.20 is identical with the angle α in Fig. 32.19 because of the similarity of the triangles in which it occurs. The angle α (see Fig. 32.19) measures the phase difference between the voltage V'_R across R and the voltage V' across the terminals AE. It determines how much sooner vector V_0 reaches the vertical position (and hence the voltage V' its maximum value) than the vector Ri_0, whose projection represents V'_R. *Since the branch current i' is in phase with the voltage V'_R, we can also say that the angle α measures the phase difference between the voltage V' and the branch current i'.*

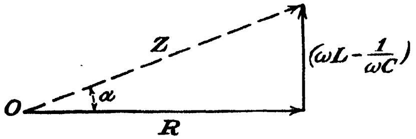


FIG. 32.20.

From Fig. 32.20 follows

$$\boxed{\tan\alpha = \frac{[\omega L - (1/\omega C)]}{R}} \tag{32.14}$$

In the special case of a circuit consisting of R and L only, $\tan\alpha = \omega L/R$ is positive, which means that the voltage V' is leading the branch current. On the other hand, in the case of a circuit consisting of R and C only, $\tan\alpha = -1/R\omega C$ is negative, which means that the voltage V' is lagging in phase behind the branch current.

The preceding equations are the general expressions for the impedance Z of a series connection of circuit elements and for the phase angle which measures the phase difference between the current and the voltage across the end points of the series circuit.

We shall now write the expression for the impedance in an RCL circuit, which is often referred to as *generalized Ohm's law for alternating current* although it has nothing to do with Ohm's law. There is only a formal analogy between the two expressions, the impedance Z being analogous to R in the expression of Ohm's law.

$$\frac{V_0}{i_0} = \frac{V}{i} = Z = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \tag{32.15}$$

or:

$$\boxed{i = \frac{V}{\sqrt{R^2 + \left[\omega L - \left(\frac{1}{\omega C}\right)\right]^2}}} \tag{32.15a}$$

Power dissipation in form of heat takes place only in those parts of the circuit for which $R \neq 0$. The heat generated per second is given by i^2R, where i is the effective current.

Series Resonance

Inspection of equation (32.13) leads to a very interesting prediction. The impedance of an RCL series circuit is seen to depend upon $\omega = 2\pi f$. It should therefore vary with frequency. *There should be a special value of frequency for which $\omega L = 1/\omega C$. At this frequency Z becomes $= R$; in other words, the circuit behaves as if the coil and the condenser were absent.* At all other frequencies $Z > R$. This phenomenon is called **resonance,** and the frequency at which the impedance Z reaches its minimum value is called the **resonance frequency** f^*. From the **resonance condition**

$$\boxed{\omega L = \frac{1}{\omega C}} \tag{32.16}$$

we get

$$\omega^2 = \frac{1}{LC} \qquad \therefore \omega = \frac{1}{\sqrt{LC}} = 2\pi f^*$$

and

$$f^* = \frac{1}{2\pi}\frac{1}{\sqrt{LC}} \tag{32.16a}$$

We see that the resonant frequency increases when L or C or both are decreased.

If $R = 0$, the impedance becomes zero at the resonant frequency. Equation (32.15a) shows the danger of this situation. Even a small a-c voltage across the circuit would tend to give rise to an infinitely large current. Hence, in many instances the resonance condition is to be avoided as a danger to the apparatus fed by the alternating current. There are, however, applications in which the adjustment of the circuit to resonance is desired. For instance, the familiar process of "tuning" a radio is an adjustment for resonance achieved by varying the capacitance of the condenser in the resonant RCL circuit.

From equation (32.14) we infer that at resonance the voltage V_{AB} (see Fig. 32.18) is in phase with the current since $\tan \alpha$ becomes zero.

Electrical resonance can be illustrated by the following simple experiment (see Fig. 32.21):

The oscillator *Osc* is a source of an alternating emf whose frequency can be varied without changing the amplitude. When the switch *Sw* is closed, L and C are "short-circuited" and the current delivered by *Osc* gives rise to an alternating potential difference only in the resistor R. This potential difference is applied to the input of a cathode-ray oscilloscope on the screen of which we obtain a wave pattern representing the voltage across R as a function of time. Since the current is proportional to the voltage drop across R, the pattern seen on the oscilloscope screen can also be interpreted as a representation of current versus time. When the frequency of the alternating current is increased, we see that the peaks move closer together without altering their height. If, however, we open the switch Sw before we increase the frequency of the oscillator output current, we see that the number of peaks on the screen increases as before but the height of the peaks does not remain constant. If $\omega L < 1/\omega C$ when we start out, the peaks grow in height as the frequency increases, reach a maximum height, and begin to decrease as we continue increasing the frequency. This observation can be explained as follows: ω is proportional to f; hence, ωL grows and $1/\omega C$ decreases as f is being increased. At a certain frequency $f = f^*$, ωL becomes equal to $1/\omega C$ and the impedance becomes a minimum. The current at this point is at a maximum, and so is the voltage drop across R. Hence, the amplitude of the waves observed on the oscilloscope screen reaches its maximum value. A further increase in f restores a finite value of the brackets $[\omega L - (1/\omega C)]$. At much higher frequencies $1/\omega C$ can be neglected against ωL, and the impedance can be approximated by ωL (neglecting R). We see clearly how in that region of frequencies the impedance grows in proportion to f, reducing the current to smaller and smaller values.

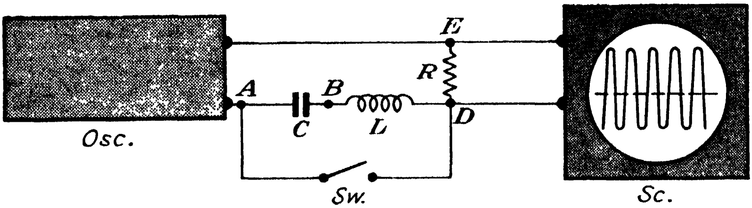


FIG. 32.21. Experimental demonstration of resonance in an a-c circuit.

Our explanation of resonance was based so far on a mathematical analysis of equation (32.13). Is it possible to give a nonmathematical physical explanation of resonance? Such an explanation can be given quite easily. It can be presented most clearly in connection with the following modification of the preceding experiment. If, instead of applying an alternating *emf* of constant amplitude to the circuit of Fig. 32.21, we send an alternating *current* of constant amplitude through it while the terminals A and E are connected to the oscilloscope, we notice that at resonance ($f = f^*$) the voltage amplitude measured by the oscilloscope becomes a minimum, being the same between A and E as between D and E, while the potential difference between A and D becomes zero. In other words, the portion AD of our circuit behaves like a zero impedance. If we connect the oscilloscope to the points A and B, we notice that $V'_{AB} \neq 0$ and similarly we show that $V_{BD} \neq 0$. So how can V'_{AD} be constantly zero? Oscilloscopic measurements show that $|V'_{AB}| = |V'_{BD}|$, but both voltages are in phase opposition! Hence, the voltages across C and L add up to zero at any moment. This is, however, something that we could have predicted from the knowledge that the voltage across the capacitance lags the branch current by 90° whereas the voltage across the inductance leads the current by 90°. Hence, the voltages across L and C must be in phase opposition, and the potential difference measured between points A and D of Fig. 32.21 represents the difference between the instantaneous voltages across L and C. In the special case of resonance the amplitudes of these two voltages are equal, and hence they add up to zero.

Example

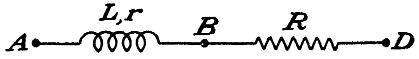
An alternating current flows through a coil and a resistor R, which are connected in series. The coil has the ohmic resistance r, and its coefficient of self-inductance is $L = 0.015$ henry. $r = 1$ ohm, $R = 3$ ohms, and we measure the following values: for the effective current, $i = 2$ amp and for the effective voltage, $V_{AD} = 10$ volts.

Find a. The impedance Z between points A and B.

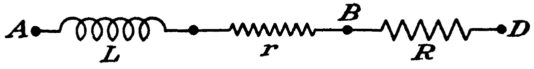
b. The frequency of the alternating current.

c. The phase lag θ between i' and V'_{AD}.

d. The effective voltage (V_{AB}) across AB.



Solution. We draw an equivalent circuit in which r is drawn as a resistor in series with L:



1. $$Z = \frac{V_{AD}}{i} = \frac{10}{2} = 5 \text{ ohms} \qquad (32.15)$$

2. $$Z = \sqrt{(R + r)^2 + (\omega L)^2} \qquad (32.13)$$

$$25 = (3 + 1)^2 + (\omega L)^2 \rightarrow (\omega L)^2 = 9; \qquad \omega L = 3 \text{ ohms}$$

$$\omega = \frac{3}{L} = \frac{3}{0.015} = 200 \text{ rad/sec}$$

Now $\omega = 2\pi f$; hence, $f = \omega/2\pi = 200/2\pi = \underline{31.8 \text{ cycles/sec}}$

3. $$\tan\theta = \frac{\omega L}{(R + r)} = \frac{3}{4}$$

4. $V_{AB} = 2Z_{AB}$; $Z_{AB} = \sqrt{r^2 + (\omega L)^2} = \sqrt{1^2 + 9} = \sqrt{10} = \underline{3.17 \text{ ohms}}$

$$V_{AB} = 2 \times (3.17) = \underline{6.34 \text{ volts}}$$

QUESTIONS AND PROBLEMS

1. How does the impedance of an RCL series circuit depend on the a-c frequency?
2. Can the impedance of a circuit ever become zero?

3. What alterations would you have to make in order to transform an a-c generator into a d-c generator?

4. A device is to be protected against a sudden *transient* increase in voltage. How could you protect it?

5. A device through which a direct current is flowing is to be protected against a possible flow of alternating current of a high frequency. How could you accomplish this?

6. The a-c current through a lamp is to be cut down without wasting an appreciable amount of energy due to generation of heat. Suggest three different ways of doing this.

7. Name an instrument which can measure instantaneous and peak values of alternating current and voltage.

8. Name an instrument which could indicate instantaneous values of alternating current and voltage. Draw a diagram indicating how you would arrange to measure and to record photographically these quantities.

9. Can you suggest what type of instrument could be adapted to the measurement of effective values of current and voltage?

10. Design an a-c circuit in such a way that it would pass a current of

$$f = 1{,}000 \text{ cycles/sec}$$

in preference to alternating currents of other frequencies (filter).

11. Design a circuit which will resonate at 60 cycles/sec and in which 120 cal of heat will be generated each minute by a current of 2 amp.

12. An alternating voltage of 110 volts effective value and 60 cycles/sec is connected to an immersion heater of 110 ohms resistance ($L = 0$). How long will it take to heat 1 liter of water from the temperature of 20°C to the boiling point?

13. A voltage of effective value $V = 100$ volts and frequency of 400 cycles/sec is connected across a series combination of a 10-ohm resistor and a 1-microfarad capacitor. Determine (*a*) the peak values of the voltage (1) across the resistor, (2) across the capacitor; (*b*) the heat generated each second in this circuit; (*c*) the impedance of the circuit; (*d*) the phase difference between the current and the applied voltage.

14. A current of effective value $i = 2$ amp flows through a coil at a frequency of 30 cycles/sec. The effective value of the voltage measured across the coil is 10 volts. The inductance of the coil is 0.015 henry. Find (*a*) the impedance of the coil, (*b*) the ohmic resistance of the coil, (*c*) the power dissipated in form of heat.

15. An *RCL* series circuit consists of a 0.1-microfarad capacitor, a 0.1-henry inductance, and a 10-ohm resistor.

a. What is the resonant frequency of the circuit?

b. What is the impedance at resonance?

c. How large are the peak voltages across L and C at resonance if the peak current is 10 amp?

d. How large is the circuit impedance at a frequency (1) 5 times as large as the resonance frequency, (2) 5 times as small as the resonance frequency?

CHAPTER 33

ELECTROMAGNETIC OSCILLATIONS AND WAVES. MAXWELL'S THEORY OF LIGHT

a. Draw a parallel-resonance circuit.

b. How can you determine the natural frequency of a parallel-resonance circuit?

c. What energy transformations take place periodically in a parallel-resonance circuit?

d. What is the phase relationship (1) between the potential drops across the branches L and C of the circuit of Fig. 33.1? (2) between the currents through these branches?

e. How large is the current flowing through the wire section Aa of the circuit of Fig. 33.1 at resonance? How do you explain your answer? Can you dispense with the generator G?

f. Why does the same formula hold for the resonance frequency in series resonance as well as in parallel resonance?

g. Can you choose any value of L and C in designing an oscillatory circuit? Describe what takes place in an oscillating LC circuit.

h. How does a spark-gap oscillator function?

i. What is a dipole oscillator?

j. Describe the electromagnetic field in the vicinity of a dipole oscillator. How does this field differ from the field at a great distance?

k. What is the relationship between the directions of the electric field vector and the magnetic field vector at a great distance from the oscillator?

l. Describe an electromagnetic wave. What is the phase relation between the electric and magnetic vector?

m. How can electromagnetic waves be generated? How can they be detected?

n. How did Hertz demonstrate the existence of electromagnetic waves? How did he determine their velocity of propagation?

o. What do we mean by the statement: "Electromagnetic waves are polarized"?

p. How can you demonstrate the polarization of electromagnetic waves?

q. What is the connection between Maxwell's equations and Fig. 33.15?

r. What was the main evidence which suggested to Maxwell that light is an electromagnetic wave motion?

s. In what way is the electromagnetic light theory superior to the mechanical ether theory of light?

t. What phenomenon was the elastic ether theory unable to explain without making unreasonable assumptions?

u. How do you picture the process of emission of light by an individual atom according to Maxwell's theory?

v. How can light be polarized? How can polarization of light be detected?

w. In what way were Maxwell's original ideas about the electromagnetic field and waves modified by his successors?

Parallel Resonance

In Chap. 32, we encountered an electrical phenomenon which bore close resemblance to a familiar mechanical phenomenon, namely, the phenomenon of resonance. In the case of mechanical resonance, application of a very small periodic force to an oscillator enabled us to obtain exceedingly large amplitudes of vibration by making the frequency of vibration of the applied force equal to the "natural frequency" of the oscillator, *i.e.*, to the frequency with which the oscillator vibrates of its own accord.

In the case of electrical resonance of an RCL series circuit, a very small periodic emf enabled us to obtain a maximal electric current at a certain frequency. Can we conclude by analogy that this *resonant frequency* of the electrical circuit is equal to its *natural frequency?* In other words, can we assume that the electric circuit can act as an electric oscillator whose natural frequency is equal to the resonant frequency? This question will be answered in the affirmative by the following consideration.

Let us consider, instead of a series circuit, a parallel LC circuit (Fig. 33.1). G is a generator providing an alternating emf of constant amplitude and variable frequency. A is a meter measuring effective current. As we increase the frequency of the emf from 0 to ∞ we make the following observations:

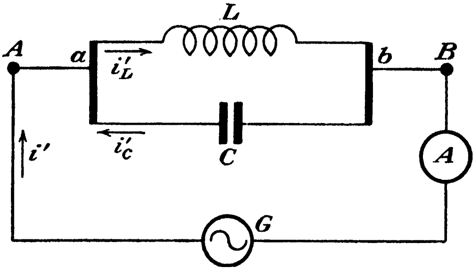


FIG. 33.1. Oscillatory circuit.

At first, the current is relatively strong, since ωL is small at low frequencies. The current flows mainly through L, since at low values of f, $1/\omega C$ is very large; *i.e.*, the capacitive impedance of C is very much larger than the inductive impedance of L. As the frequency increases, the current goes down and becomes zero at a certain frequency f^*, which happens to be exactly the frequency at which the given condenser and inductance coil resonate when connected in series. In other words,

$$f^* = \frac{1}{2\pi}\frac{1}{\sqrt{LC}} \tag{32.21a}$$

(f^* is given in cycles per second when L is measured in henrys and C in farads.)

When the frequency is increased further, the current increases again. Thus we see that our *parallel circuit is also capable of resonance and that* *the resonant frequency is the same as for a series circuit.* There is, however, an important difference between these two cases. *Whereas the impedance of a series LC circuit becomes zero at resonance, the impedance of a parallel LC circuit is seen to become infinite at resonance.*

How can this difference in behavior be explained? The explanation follows the same line of reasoning as has been used before in the discussion of parallel resonance. The wires connecting the terminals of L to those of C and to the points A and B can be considered to have approximately zero impedance at all frequencies. Hence, the potential difference across L is equal to that across C at any moment. Now we know that the current through L *lags* behind the applied voltage by 90° whereas the current through C *leads* the applied potential difference by 90°; hence, *there is a phase difference of* 180° *between the currents in L and in C; i.e., these currents are in phase opposition.* Therefore, the amplitude of the main current i' feeding the two branches is equal to the difference between the amplitudes of the branch currents: $i_0 = |(i_L)_0| - |(i_C)_0|$. The branch currents are given by the "generalized Ohm's law":

$$(i_L)_0 = \frac{V_0}{\omega L} \qquad \text{and} \qquad (i_C)_0 = \frac{V_0}{1/\omega C}$$

where V_0 is the amplitude of the common potential difference across L and C.

Now, at resonance $\omega L = 1/\omega C$; hence $|(i_L)_0| = |(i_C)_0|$, and therefore $i_0 = (i_L)_0 - (i_C)_0 = 0$.

Thus we see that at resonance very heavy currents may flow in the individual branches L and C while the current supplied by the generator to the parallel LC circuit through the leads Aa and Bb is zero as a consequence of the fact that the branch currents are equal and are in phase opposition.

Electromagnetic Oscillations

But if no current is supplied to the resonant parallel LC circuit by the generator, why not remove it altogether and with it all the wires leading from it to L and C? The inductance coil (assumed ideally of zero resistance) and the condenser should remain unaffected by the removal of the useless generator, and the heavy currents should continue flowing in the two branches of our parallel circuit happily ever after.† It should be possible to perform such an experiment using a superconductive lead coil cooled to a few degrees above absolute zero. Such oscillations of current occur also at room temperature with nonsuperconductive coils, but they persist only for a fraction of a second owing to rapid conversion of electrical energy into heat.

---

† There will be a gradual diminution of the alternating current due to electromagnetic energy radiation from the oscillatory circuit (see p. 643).

The parallel LC circuit has all the characteristics of an oscillator, and the flow of alternating currents through an isolated circuit of this type can be regarded as an electromagnetic oscillation. *The characteristic feature of any oscillator is the periodic conversion of one form of energy into another.* It can easily be shown that *in this electrical example we have a periodic interconversion of the energy of the magnetic field of the coil L into the electrostatic energy of the charged condenser C.* Let us begin our consideration at a moment when the current in both C and L is zero while the potential difference across C is a maximum, the right plate being negative. C now begins to discharge through L, giving rise to a growing current in coil L and hence to a growing magnetic field. In this process the electrostatic potential energy of C decreases as the magnetic field energy of L increases. The energy of L cannot continue to increase beyond the point where C has lost its charge, *i.e.*, when the potential difference across C (and hence also across L) has become zero. At this moment the electrostatic potential energy of the condenser is zero and the magnetic field of coil L (and hence the current through L) is at a maximum. The energy which was originally stored in our oscillator in the form of electrostatic potential energy of the capacitator is now stored in the form of the energy of the magnetic field of the coil L.

Now the magnetic field begins to collapse, and as a result, an emf is self-induced in L which, according to Lenz's rule, is such as to continue driving the current in the same sense; *i.e.*, the electrons continue moving counterclockwise. The condenser is now being charged so that the left plate is acquiring the negative charge. It is gaining potential energy at the expense of the energy of the vanishing magnetic field. This process can continue only until the condenser has acquired its original electrostatic energy, which leaves no energy for the magnetic field. The current has again dropped to zero, and the condenser discharge can again renew its cycle, but now the current flows through L in the clockwise sense. We see then how the condenser is being charged alternately by electrons which oscillate back and forth in the coil L while magnetic field energy is periodically converted into electrostatic potential energy, just as kinetic energy is periodically transformed into potential energy (and vice versa) in the case of a pendulum.†

We can now understand why the parallel LC circuit is referred to as an **oscillatory circuit.** This discussion also explains why the natural frequency of the oscillatory circuit is identical with its resonance frequency.

---

† The above description shows that there is a 90° phase displacement between the alternating electric and magnetic field, just as there is a 90° phase difference between the displacement and velocity of the pendulum.

The Spark-gap Oscillator

The preceding discussion suggests a simple design of an oscillator, *i.e.*, of a generator of an alternating emf. A condenser C connected in series with coil L and a switch Sw (Fig. 33.2) are theoretically all we need. To start oscillations in this circuit, we have merely to charge the condenser and subsequently close the switch Sw. This will initiate the oscillatory flow of electricity described above. If we connect the terminals of the coil L across an oscilloscope, however, we notice that the oscillation thus engendered does not persist for more than a few periods. *The oscillation is damped* as the record of V'_L versus time indicates (Fig. 33.3). This damping is due to the transformation of electrical energy into heat in the ohmic resistance of the circuit. To maintain oscillations in this circuit, we could charge the condenser, then close and subsequently open the switch, repeating the whole operation cyclically in rapid succession. This would maintain in our circuit damped oscillations whose character is indicated by the plot of V'_L as a function of time (Fig. 33.4). The desired effect can be obtained without the laborious process of mechanical switching by using a "spark gap" as a switch

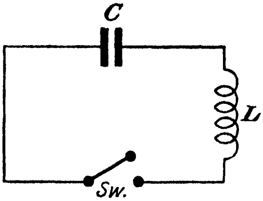


Fig. 33.2.

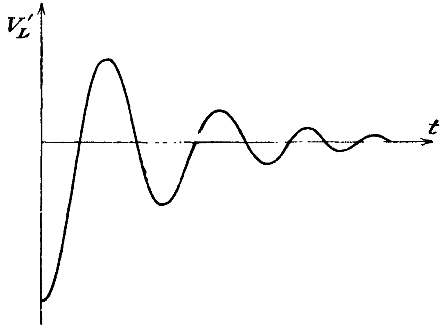


Fig. 33.3. A damped oscillation.

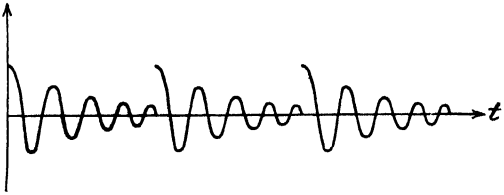


Fig. 33.4. Damped oscillations generated by a spark-gap oscillator.

as illustrated in Fig. 33.5. *P* and *S* are the primary and secondary coils of an iron-core step-up transformer connected to the 60-cycle supply line. The output of the secondary coil is applied to an oscillatory circuit which is identical with the circuit shown in Fig. 33.2 except that the switch *Sw* is replaced by a spark gap *G*, *i.e.*, by two closely placed electrodes between which a spark occurs whenever the potential difference across the gap reaches a critical value. The spark is very rich in gas ions, and hence the gap *G* remains conductive during the brief period of persistence of the spark. The "jumping" of the spark is thus equivalent to the closing of the switch *Sw*.

The mode of operation of this oscillatory circuit can then be described as follows in a simplified fashion: The peak voltage of the output of coil *S* is higher than the voltage V^* at which a spark jumps across the gap *G*. As the instantaneous voltage V'_S grows, the condenser is being charged. As soon as $V'_S = V^*$, a spark jumps across *G* and the condenser discharges through coil *L*, *G* acting now as a closed switch. As V'_S drops, the spark is discontinued until it jumps across *G* again when *C* has been charged again with reversed polarity. Thus we should expect at least 120 sparks a second with an applied a-c voltage of 60 cycles/sec. The high-frequency alternating current engendered in the oscillator is confined mainly to the oscillatory circuit, since the coil *S* connected across the gap has a very high impedance at the (usually high) frequency of the oscillations generated in the *LC* circuit.

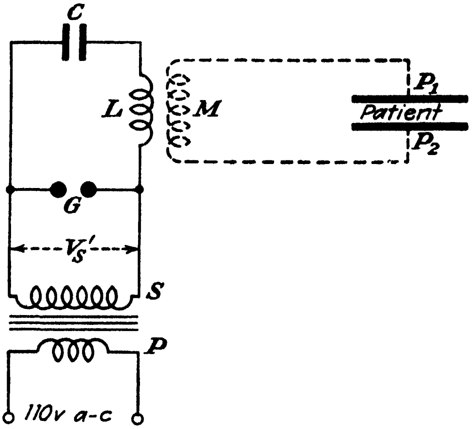


FIG. 33.5. Scheme of a spark-gap oscillator.

Such **spark-gap oscillators** were extensively used in the early days of wireless telegraphy as transmitters (see Appendix) and are still occasionally used for medical **diathermy.** In many pathological conditions it is desirable to heat a deep-seated part of the human body. Owing to the high resistance of the skin and of the superficial fat as well as to the danger of electric shock, one cannot pass a direct current or a low-frequency alternating current through the body in order to heat it. It is, however, quite safe to pass currents of several amperes through the body at frequencies above 300,000 cycles/sec. In this case, the skin and the superficial fat layers act as the dielectric of a condenser which possesses a low capacitive impedance at high frequencies and passes the alternating current readily. Figure 33.5 shows a coil M coupled inductively to coil L. The emf induced in M is applied to the two metal plates P_1 and P_2, which are placed against the skin of the patient. At very high frequencies one can afford to refrain from direct metallic contact between the plates P_1 and P_2 and the patient's body. The air gap between the plates and the patient simply acts as the dielectric of a condenser inserted in series with the circuit, offering a low capacitive impedance at the very high frequency used (around 50 million cycles per second for short-wave diathermy). At such high frequencies the magnetic field of the coil L can also be utilized directly to generate eddy currents by electromagnetic induction in the patient's body (principle of the Inductotherm machine).

The Dipole Oscillator

The order of magnitude of the frequencies of oscillation which can be obtained with an electric oscillatory circuit can be estimated easily with the aid of equation (32.21a). Assume, for instance, a capacity of $C = 10^{-10}$ farad and an inductance of $L = 10^{-2}$ henry. Then

$$f = \frac{1}{2\pi} \cdot \frac{1}{\sqrt{LC}} = \frac{1}{2\pi \sqrt{10^{-10} \times 10^{-2}}} = \frac{10^6}{2\pi} = 159{,}000 \text{ cycles/sec}$$

Very much higher frequencies of oscillation can be achieved without much difficulty. We see that, to increase f, we have to diminish L or C or both. Extremely low values of L as well as of C can be obtained by the design of the oscillator shown in Fig. 33.6, which pictures the so-called **dipole oscillator.** It is analogous to the oscillator shown in Fig. 33.5. T is a transformer, and G is a spark gap. The main difference between the two circuits consists in the design of the resonant circuit. The capacity of the circuit is given by the very small electrostatic capacity of the thin rods A_1 and A_2; the self-inductance of this circuit is also exceedingly small, being the inductance of the straight wires A_1 and

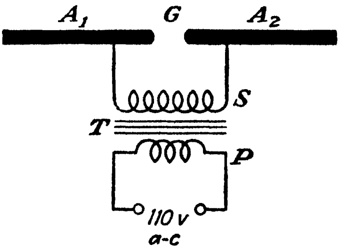


FIG. 33.6. Dipole oscillator.

A_2. Frequencies of hundreds of millions of cycles per second can be obtained with oscillators of this type. The picture of the a-c flow is particularly simple in this case. The electrons simply move back and forth inside the straight rods A_1 and A_2. Their oscillatory motion is accompanied by an alternating magnetic field whose lines of force surround the wire as shown in Fig. 23.2 for the case of a unidirectional current. At the same time, the electric polarity of the rods A_1 and A_2 alternates just as the polarity of the plates of the condenser C of the circuit of Fig. 33.5. As a result of the potential difference between A_1 and A_2 an electric field exists in the vicinity of these rods which is indicated by dashed lines in Fig. 33.7. The magnetic lines are shown as solid circles in the same diagram. In the vicinity of the dipole oscillator the electric and magnetic fields are 90° out of phase, but experiment and theory (which is too advanced for discussion in this text) indicate that, *at great distances from the oscillator, the detected alternating electric and magnetic fields are in phase and that the electric and magnetic field vectors are perpendicular to each other and to the line joining the point of observation to the oscillator.* In Fig. 33.7, P is a point of observation located along the X axis. The magnetic field line, which passes through P (but is not drawn in Fig. 33.7), is a circle in the XZ plane. The electric line of force at point P intersects this circle at right angles. The two vectors E and H, representing the instantaneous electric and magnetic field intensities at P, indicate that these fields are perpendicular to the X axis which joins P to the center of the oscillator.

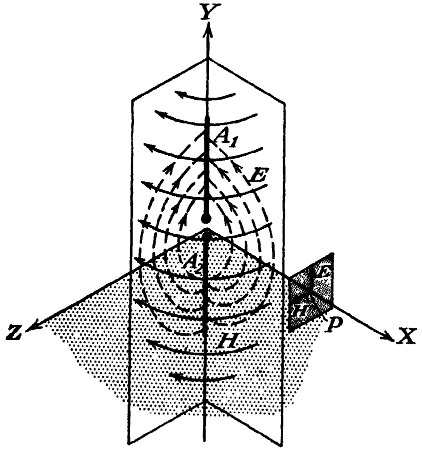


FIG. 33.7. The electric and magnetic fields near the dipole oscillator. (*After Sears and Zemansky, "College Physics," Addison Wesley.*)

Electromagnetic Resonance

Let us now consider the following experiment: Two rods B_1 and B_2 (Fig. 33.8), similar and parallel to the rods A_1 and A_2 of the setup of Fig. 33.6, are placed at point P of Fig. 33.7, while A_1 and A_2 are connected to a transformer T as shown in Fig. 33.6, so that the arrangement functions as a dipole oscillator. The length of the rods B_1 and B_2 can be varied in telescope fashion as is apparent from Fig. 33.8. At a certain length of these rods, tiny sparks begin to jump across the narrow gap between B_1 and B_2. Under these conditions the value of the capacitance and self-inductance of the arrangement consisting of B_1 and B_2 is such that the expression $1/\sqrt{LC}$ is the same for the "dipole antenna" (A_1A_2) as for the dipole antenna (B_1B_2). In other words, both systems have the same natural frequency; *they are in resonance.* The oscillations which take place in the dipole oscillator (A_1A_2) evidently excite similar oscillations in (B_1B_2), which, at resonance, generate a high enough voltage amplitude between B_1 and B_2 to cause a spark to bridge the gap. This spark indicates that a considerable amount of energy is converted into light, sound, and heat. Where does this energy come from? The only possible source of energy is the oscillator (A_1A_2). How is the energy transferred from (A_1A_2) to (B_1B_2)?

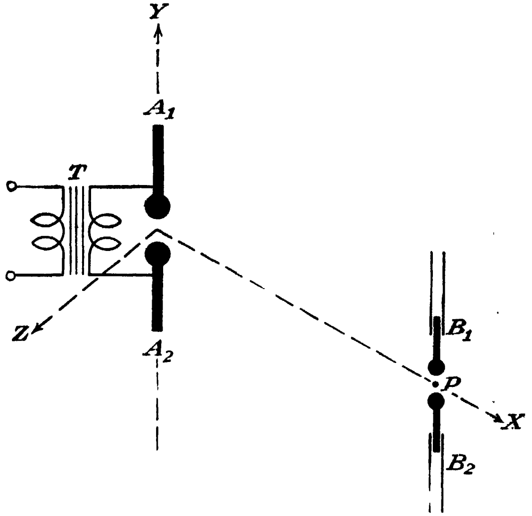


FIG. 33.8.

The energy transfer from oscillator (A_1A_2) to the "tuned" oscillator (B_1B_2) bears a strong resemblance to the experiment on acoustical resonance illustrated in Fig. 27.17. The air molecules are caused to vibrate in the adjustable air column by periodic forces exerted upon the air in the cylinder by the passing rarefactions and condensations of the sound wave emitted by the tuning fork. Can we infer from the analogous generation of oscillations of the electrons in the antenna (B_1B_2) that an electric wave, consisting of electric fields of alternating strength and direction, is emitted by the antenna (A_1A_2) and that this alternating electric field shakes the electrons up and down as the wave passes the antenna (B_1B_2)? Maxwell predicted that an oscillatory circuit, such as the circuit shown in Fig. 33.6, would emit not only an electric wave but also an analogous magnetic wave, the two waves being inseparable and forming a so-called *electromagnetic wave.* Maxwell's prediction was confirmed by the brilliant experiments of Heinrich Hertz. To facilitate comprehension, we shall deviate from the historical sequence and discuss Hertz's experiments before Maxwell's theory.

Electromagnetic Waves

Heinrich Hertz was the first to demonstrate experimentally the energy transfer illustrated in Fig. 33.8 and to show that a wave motion is involved in this process. Just as we have established the wave nature of sound and of light by producing the characteristic interference patterns of standing sound or light waves, Hertz used as an argument for the wave nature of the electromagnetic energy transfer the results of the following experiment, which he interpreted as a demonstration of standing electromagnetic waves.

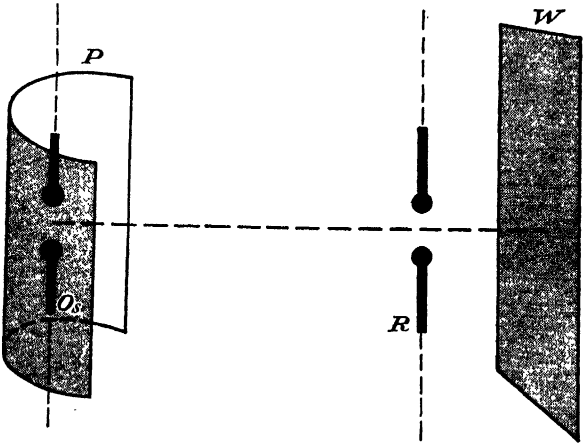


FIG. 33.9. Hertz's experiment on generation of standing electromagnetic waves.

Hertz placed a dipole oscillator *Os* along the focal line of a parabolic metal mirror *P* (Fig. 33.9).† The oscillator *Os* was placed about 2 m from the wall *W*, which was covered with zinc foil. The *resonator R* could be moved along the perpendicular dropped from the spark gap of *Os* upon the wall *W*. When the oscillator *Os* was "excited," tiny sparks could be seen to jump across the gap of the resonator *R*, *but only at*

---

† If *Os* were a linear source of sound or of light, the reflection of these waves from this mirror would provide us with a parallel beam of sound or light waves (provided the wave length were small in comparison to the dimensions of the mirror). Similarly, we obtain a parallel beam of electromagnetic waves.

certain distances of R from the wall. Hertz described his observations as follows:

> . . . in the immediate vicinity of the zinc wall no sparks are seen to jump across the resonator gap. When the resonator is moved gradually from the wall they appear and reach a maximum and decrease again as the resonator is moved farther and farther from the wall in order to disappear again at a certain point. As the resonator is moved still farther we reach a second maximum, then again a minimum, etc. . . . .
>
> The distance between the points O at which the sparks vanish turns out to depend on the size of the oscillators as well as on their periods of oscillation. The diagram [Fig. 33.10] illustrates the result of a specific experiment.
>
> These phenomena show the characteristic feature of standing waves of the electric stress whose nodal points are given by the minima and whose anti-nodes by the maxima of the lengths of the sparks. These interference phenomena can be explained only by the reflection from the metal wall (W) of a wave train emanating from the primary circuit (oscillator) into the surrounding air space—any other interpretation is ruled out.

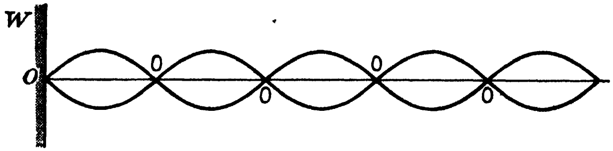


FIG. 33.10. Standing-wave pattern obtained in Hertz's experiment.

This experiment enabled Hertz to determine the velocity of the assumed electromagnetic waves. According to his interpretation, the distance between two adjacent minima should be equal to $\lambda/2$, where λ is the wave length of the electromagnetic wave train (compare the treatment of standing waves in Chap. 27). The frequency of the oscillator was computed from the known capacity and self-inductance of the dipole arrangement. Thus Hertz was able to calculate the velocity c from the relation $c = f\lambda$. He obtained very nearly $c = 3 \times 10^{10}$ cm/sec, which, as we recall, is equal to the velocity of light in a vacuum in accordance with Maxwell's prediction.

A glance at Fig. 33.7 suggests that an observer stationed at point P will observe an electric field representable by lines of force intersecting the X axis at P at right angles to the XZ plane. Their direction is indicated by vector E. In other words, the electric field at P is parallel to the antenna. On the other hand, when we consider a circular magnetic line of force passing through point P, we see that it is perpendicular to the X axis and to the E vector, its direction being indicated in the diagram by the vector H. A theoretical consideration far beyond the scope of this book shows that, *no matter where the point P is located, the magnetic vector H is perpendicular to the antenna and to the line joining P*

to the center of the spark gap whereas the electric vector E is perpendicular to the latter line and to the magnetic field vector. The oscillation of the electric field intensity is in phase with the oscillation of the magnetic field intensity (provided the distance of P from the antenna is very large as compared with the wave length).

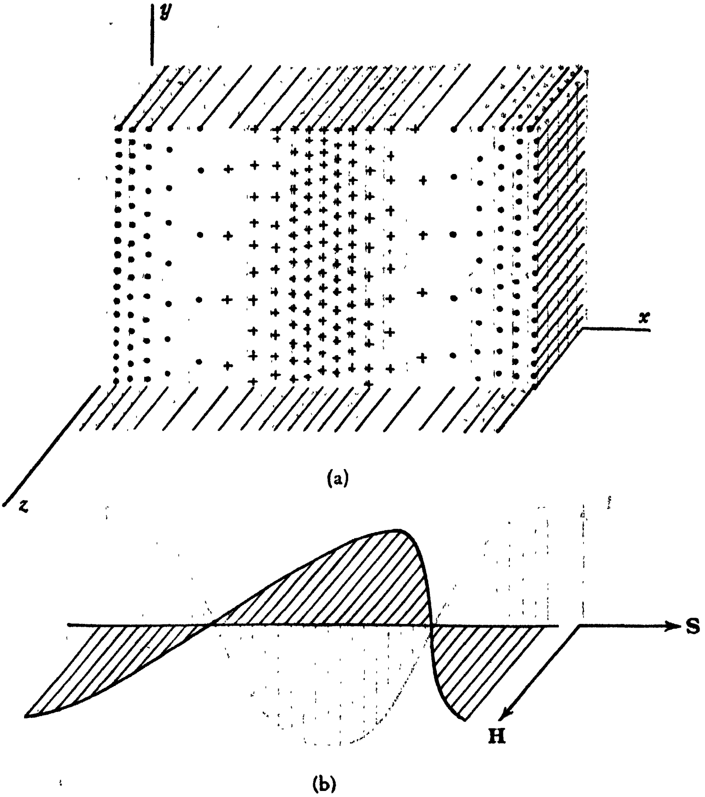


Fig. 33.11*a*. Electric (red) and magnetic (black) lines of force in a progressing plano electromagnetic wave. *b*. The distribution of the electric (red) and magnetic (black) field intensity in a plane electromagnetic wave. (*From Emery, "Ultra-High-Frequency Radio Engineering,"* Macmillan.)

Figure 33.11*a* illustrates the relation between the electric and magnetic fields at large distance from P, where the curvature of the waves is so small that they can be considered as *plane waves.* The electric lines of force are represented by red lines and the magnetic ones by black lines. Crosses are the tails of vectors pointing away from the reader, and dots are the points of vectors aiming toward the reader. The wave travels in the direction of the positive X axis. The change in the intensity of the fields is represented by the variation in the density of the lines of force, and the reversal of the direction of the field is indicated by a reversal of the arrows.

Figure 33.11*b* shows the electric and magnetic wave in the conventional transverse representation. The red curve (in the vertical plane) shows the electric field distributions along the X axis, and the black curve (in the horizontal plane) shows the magnetic field. The crests of both waves are seen to coincide in space.

The above considerations show that the electromagnetic waves are what we defined as *polarized waves*. They display an asymmetry when viewed along the direction of propagation. In the case illustrated above, for instance, electrons would be made to oscillate along the vertical but not along the horizontal axis.

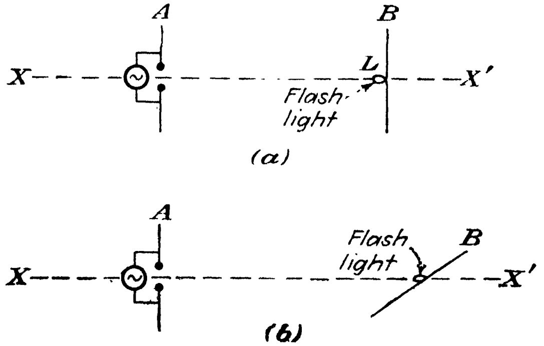


Fig. 33.12. Demonstration of the polarization of electromagnetic waves. *a*. The receiving antenna is parallel to the transmitting antenna. *b*. The receiving antenna is perpendicular to the transmitting antenna.

The polarization of electromagnetic waves can be easily demonstrated by the following experiments:

Instead of using the Hertzian spark-gap resonator (B_1B_2 of Fig. 33.8), we bridge the gap between B_1 and B_2 by a small flashlight bulb. Electromagnetic oscillations in this "receiving antenna" are then indicated by a glow of the bulb. If the receiving antenna B is held parallel to the transmitting antenna (oscillator A of Figure 33.12*a*), the electric vector of the waves which reach B is parallel to B. Electrons in B are agitated back and forth by the alternating electric field of the passing wave, so that an alternating current flows through the bulb and lights it.

If we now rotate the receiving antenna B about the axis X-X', the light gets dimmer and is finally extinguished when the antenna B is perpendicular to A (Fig. 33.12*b*).

Polarization of the waves emitted by the antenna A can also be demonstrated in the following fashion (Fig. 33.13). A wire grid G is interposed between the transmitting antenna A and the receiving antenna B. The lamp L of antenna B is lit as long as the wires of the grid G are oriented horizontally. As the grid is rotated about the axis X-X', the lamp L glows dimmer and dimmer until it is extinguished when the wires of G have been oriented vertically, *i.e.*, parallel to the antennas A and B. The effect is most pronounced if the length of the wires of the grid is equal to half the wave length of the electromagnetic waves emitted by A. In this case the wires resonate at the frequency of the arriving waves. Electromagnetic vibrations are set up in the wires such that the center of the wire is a potential node and the ends are potential antinodes. Each of the wires now acts as a transmitting antenna, reradiating electromagnetic waves into space. We shall call the wave radiated by G in the direction back toward A the reflected wave.† This wave interferes with the original wave coming from A and gives rise to the interference phenomena described by Hertz (see page 645). The wave emitted toward B also interferes with the original wave (which is emitted by A and passes through G). The result of our experiment (extinction of the lamp L) shows that the wave reradiated by G is 180° out of phase with respect to the original wave, which passes through G, so that the effects of the two waves upon B are mutually canceled. A cancellation effect is not observed, however, when the wires of G are horizontal, because no appreciable electron oscillations are induced in them by the incident wave in this case in which the electric vector is perpendicular to the wires. The consideration of reradiation of electromagnetic waves by antennas and of polarization will play an important role in subsequent discussions.

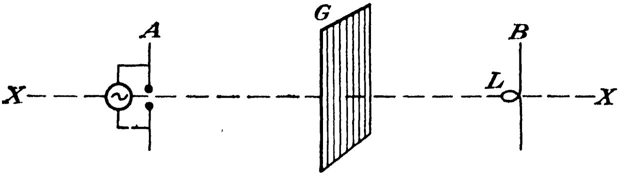


FIG. 33.13. Extinction of polarized electromagnetic waves by a wire grid.

Maxwell's Prediction of Electromagnetic Waves

The discovery of electromagnetic waves was for more than one reason a memorable event in the history of physics. The fact that the existence of electromagnetic waves was predicted by Maxwell and that their properties, such as the velocity of propagation in various media, were correctly described by him more than 20 years before their actual observation by Hertz constitutes one of the greatest triumphs of abstract theoretical reasoning. The fact that the attention of the physicists was focused not on interacting particles but on the conditions in the space around them heralded a new way of physical thinking, a transition from particle theories to field theories. Maxwell's great contribution consisted largely of a mathematical formulation and extension of Faraday's ideas on the nature of electric and magnetic fields.

---

† Such grids as G are frequently used as "tuned" mirrors for electromagnetic waves.

To Faraday the electric and magnetic lines of force were not merely a means of a convenient geometrical representation of a vector field. Faraday's lines of force were a kind of strings, each having a tendency to contract and to repel neighboring lines of force. He considered them as the essential agents of electromagnetic interactions, the charges and poles being merely the end points of the lines of force. Faraday's lines of force consisted of chains of polarized molecules as shown in Fig. 33.14. Even in a vacuum he imagined polarizable particles to be present, which we shall call "ether molecules," behaving in a similar fashion as an emulsion of mercury droplets in oil would behave. In an electric field, the positive and negative charges in such droplets would be separated by electrostatic induction, as illustrated in Fig. 15.9*a*. The droplets, or ether molecules, are thus transformed into electric dipoles which form straight chains in Fig. 15.9*a* and curved chains in Fig. 33.14. The same picture holds also for polarizable molecules of dielectrics whose polarization is superimposed upon the polarization of the ether. A similar picture of magnetic polarization of ether molecules was used to picture the structure of magnetic lines of force. It was this state of polarization of ether molecules which was considered by Faraday as the essence of an electric and a magnetic field.

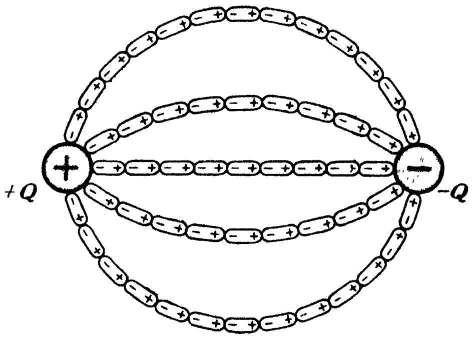


FIG. 33.14.

This model of the ether suggested to Maxwell the introduction of a new concept, the concept of **displacement current,** which plays a fundamental role in Maxwell's electromagnetic theory. Let us imagine the condenser plates of Fig. 15.9*a* to be separated by empty space. Thus the dipoles shown in the diagram are to be interpreted as polarized ether molecules. When the electric charge on the condenser plates is zero, the ether molecules are unpolarized. When the charge on the condenser plates and hence the electric field between them increases, the opposite charges in the individual ether molecules are moving apart and continue to move apart as long as the electric field continues to grow. In this process, as illustrated by Fig. 15.9*a*, the positive charges are displaced downward and constitute a positive downward current. The negative charges which are displaced upward constitute an upward negative current which is equivalent to a positive downward "displacement current." *Thus according to Maxwell, the condenser does not constitute an interruption of an a-c circuit. A displacement current flows through it as long as the intensity of the electric field between the plates changes, even if the plates are separated by a vacuum.*† Maxwell predicted that the displacement current in a vacuum would have all the characteristics of an electric current, including a magnetic field. His assertions have been eventually confirmed by experiment.

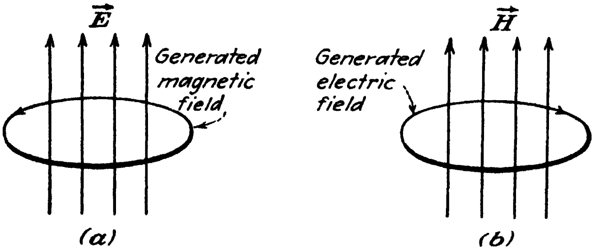


FIG. 33.15.

The hypothesis of displacement currents flowing in ether molecules makes it easy to visualize the magnetic field generated in a vacuum in the presence of a changing electric field. In the above example, the *positive* charges moving *downward* within each ether molecule in Fig. 15.9*a* are surrounded by circular magnetic lines of force which appear clockwise when viewed in the direction of motion. The *negative* charges moving *upward* are also surrounded by circular magnetic lines of force having the same sense of rotation. Thus, we see that **a changing electric field gives rise to a circular magnetic field similar to the magnetic field generated by an electric current** (Fig. 33.15*a*).

This law is the analogue of Faraday's law of electromagnetic induction (Chap. 14) which implies that a **changing magnetic field gives rise to**

---

† The assumption of polarized ether molecules is unnecessary for the definition of a displacement current. The essential manifestation of a displacement current is the variation of the electric field intensity whose time rate of change is a measure of the displacement current.

circular electric lines of force similar to the magnetic lines of force which surround a current (Fig. 33.15*b*).

The above statements are the fundamental laws of the electromagnetic field in a qualitative formulation. From his mathematical formulation of the laws of the electromagnetic field, Maxwell deduced the following relations:

$$\boxed{\ddot{E} = \frac{c^2}{\mu K} E''} \tag{33.1}$$

and

$$\boxed{\ddot{H} = \frac{c^2}{\mu K} H''} \tag{33.2}$$

In these equations, E'' stands for the "space rate of change of the space rate of change" of the electric field intensity E and $\ddot{E}$ for the "time rate of change of the time rate of change" of E. The meanings of H'' and $\ddot{H}$ are analogous to E'' and $\ddot{E}$. μ is the magnetic permeability and K the dielectric constant of the medium considered, and c is the ratio of the electromagnetic to the electrostatic unit of charge. We shall now proceed to clarify the meaning of these terms and to show how the above equations suggested to Maxwell the existence of electromagnetic waves.

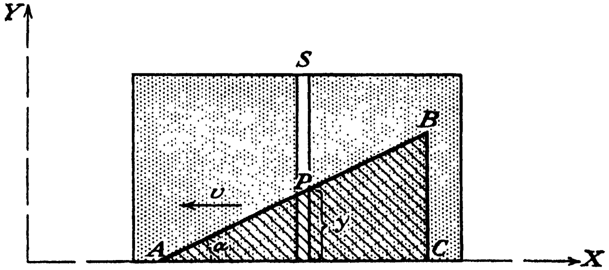


FIG. 33.16.

Let us begin with the following preliminary consideration. A triangle ABC is moving along the X axis behind a vertical slit S (Fig. 33.16). A portion P of the line AB will be visible through the slit at the height y. As the triangle moves to the left, the point P will be seen to move upward. The upward velocity of P—the "time rate of change" of y, designated by $\dot{y}$—depends on two factors: (1) the speed v of the triangle and (2) the slope y' of the line AB which we designate as the "space rate of change" of y: $y' = \tan \alpha = BC/AC$. The steeper the slope the greater is $\dot{y}$ at a given speed v. We express this proportionality of $\dot{y}$ to y' and v as follows:

$$\dot{y} = vy' \tag{1}$$

Now let us consider the acceleration of the point P, or the "time rate of change of the time rate of change" of y, which we shall designate by $\ddot{y}$. If v equals a constant, the motion of point P along the slit is not accelerated as long as AB is a straight line. As soon as the line AB is replaced by a curve (Fig. 33.17), the upward motion of P becomes accelerated because the slope y' of the portion which is seen through the slit constantly increases [see equation (1)]. We obtain an expression for the acceleration of the vertical motion of point P as follows: Consider the curvilinear triangle of Fig. 33.17 to move for Δt seconds to the left. The distance traversed will be $\Delta x = v\Delta t$. During this time interval, the upward velocity of P changes from the value $\dot{y}_1$ to $\dot{y}_2$. Hence, the vertical acceleration of P is given by the definition equation

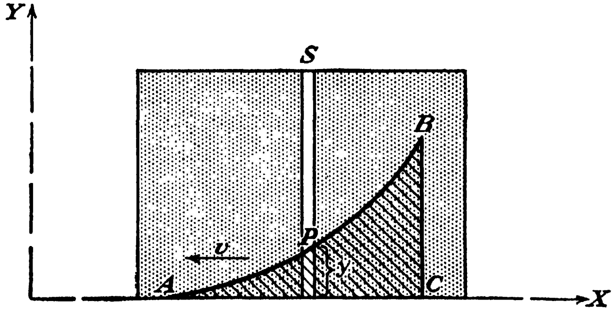


Fig. 33.17.

$$\ddot{y} = \frac{\dot{y}_2 - \dot{y}_1}{\Delta t} \tag{2}$$

and since $\Delta t = \Delta x/v$, we can write

$$\ddot{y} = \frac{\dot{y}_2 - \dot{y}_1}{\Delta t} = \frac{\dot{y}_2 - \dot{y}_1}{\Delta x/v} = v\,\frac{\dot{y}_2 - \dot{y}_1}{\Delta x} \tag{3}$$

Substituting the expression for $\dot{y}$ from equation (1) into equation (3) we obtain

$$\ddot{y} = v\,\frac{vy'_2 - vy'_1}{\Delta x} = v^2\,\frac{y'_2 - y'_1}{\Delta x} \tag{4}$$

The quotient $(y'_2 - y'_1)/\Delta x$ measures the change of slope per unit distance (or the "space rate of change of the space rate of change" of y). We shall represent it by the symbol y'' and write equation (4) as follows:

$$\boxed{\ddot{y} = v^2 y''} \tag{33.3}$$

We made no special assumptions about the shape of the curved line AB in Fig. 33.17. We could have just as well assumed the shape shown in Fig. 33.18 without changing our results.

The curve AB shown in Fig. 33.18 does not have to be imagined as a wire bent into a wave shape. It can actually represent the contour of a moving wave. The up-and-down motion of the particles of the medium through which the wave moves must then satisfy equation (33.3), which we shall call the **fundamental wave equation.**

The close analogy between his electromagnetic equations, (33.1) and (33.2), and the fundamental wave equation, (33.3), suggested to Maxwell the idea that electric and magnetic fields could be propagated in media in the form of waves in a fashion similar to the displacement y. He identified the factor $c^2/\mu K$ of his equations with the square of the wave velocity v^2 of equation (33.3) and predicted that such electromagnetic waves would travel in material media with the speed

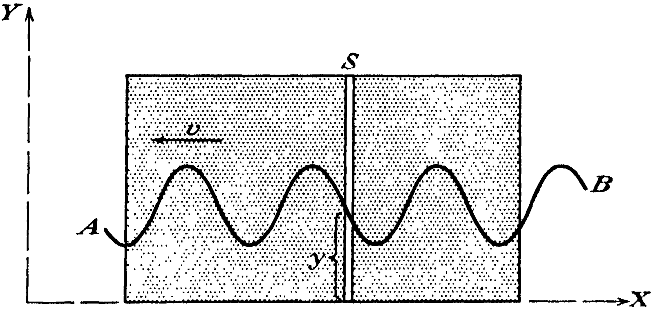


Fig. 33.18.

$$v = \frac{c}{\sqrt{\mu K}} \tag{33.4}$$

and in a vacuum, where $\mu = K = 1$, with the speed $v = c$. Since the ratio of the electromagnetic and electrostatic units of charge is $c = 3 \times 10^{10}$ cm/sec, he predicted that electromagnetic waves would travel in a vacuum with the speed of light.

Maxwell's prediction of the existence of electromagnetic waves is an excellent example of the successful use of analogical reasoning in theoretical physics.

The identity of the velocity of the electromagnetic waves with that of light suggested to Maxwell the assumption that light itself is an electromagnetic wave motion. But is it justifiable to adopt a hypothesis merely to explain one fact, namely, the coincidence of the velocity value for two wave motions? Use of such hypotheses (*ad hoc* assumptions) is considered bad scientific practice. The justification of a hypothesis lies in its ability to correlate a great number of known phenomena and to lead to a prediction of new ones. But Maxwell's theory turned out to be not an *ad hoc* hypothesis but rather one of the most fruitful physical theories, which transformed the field of optics into a branch of electromagnetism. It came as a salvation at a time of great need and confusion in the field of optics when the elastic ether theory of light was doomed to collapse.

In order to appreciate the significance of Maxwell's theory of light, we shall review briefly the status of the light theory prior to Maxwell's contribution and discuss the phenomenon of polarization of light which proved an insurmountable hurdle for the pre-Maxwellian theories of light.

The Crisis of the Mechanical Wave Theory of Light

In Chaps. 28 and 29, we discussed various arguments, some of which favored the wave theory of light and others of which supported the corpuscular theory. The wave theory proved superior to the corpuscular theory in offering explanations of the phenomena considered so far with a smaller number of basic assumptions. There was, however, a phenomenon known already in Newton's time which the wave theory seemed powerless to explain, namely, the polarization of light. Since this phenomenon is basic to the argument of this chapter, we shall pause for a page to review it before we resume the theoretical considerations.

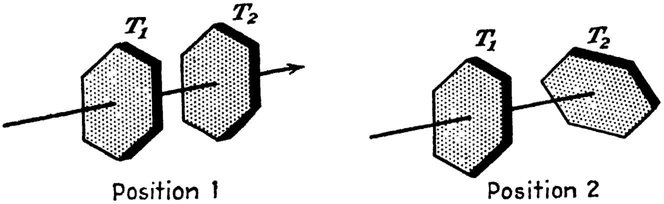


FIG. 33.19. Polarization of light by tourmaline.

When light is passed through a tourmaline crystal T_1 (Fig. 33.19), it acquires a peculiar "sidedness." A second crystal of tourmaline, T_2, behaves either as a transparent object (position 1) or as an opaque object (position 2) with respect to the beam of light transmitted by T_1, depending on its orientation. When the second crystal is oriented parallel to the first, the light passes through it. When it is rotated through 90° about the beam of light as an axis, it absorbs the light transmitted by T_1. The beam of light which passed through T_1 does not behave like light coming directly from the source of light. If T_1 is removed, T_2 transmits light equally well no matter how it is rotated about the incoming beam of light.

The light transformed by passage through the crystal T_1 is called **polarized light.** The crystal T_1 is called the *polarizer.* The crystal T_2, which enables us to demonstrate the asymmetrical behavior of the light transmitted through T_1, is called the *analyzer.* Instead of using large tourmaline crystals, one can use microscopically small crystals of iodosulfate of quinine, which can be oriented parallel to each other by a special process and mounted on a film of nitrocellulose. Such films behave like tourmaline crystals. They are commercially available under the name of *polaroid.*

About 1800 Malus discovered that light can also be polarized by reflection. Figure 33.20 illustrates an experiment which demonstrates polarization by reflection. If a beam of light is reflected from a glass plate at an angle of incidence of 57°, the reflected ray is transmitted by tourmaline T, oriented as shown in position 1 and completely absorbed by the tourmaline crystal when the latter is oriented as shown in position 2. As the crystal is turned through 90° from position 1 to 2, the intensity of the transmitted light diminishes gradually toward zero. If the angle of incidence is different from 57°, the intensity of the beam transmitted by T in position 2 is not zero. The light is said to be *incompletely polarized* in this case. The angle of incidence for which the polarization of the reflected ray is *complete* is called the *polarizing angle.* Brewster discovered that the polarizing angle p for a given substance is that angle of incidence for which the reflected ray R and the refracted ray T are perpendicular to each other (Fig. 33.21).

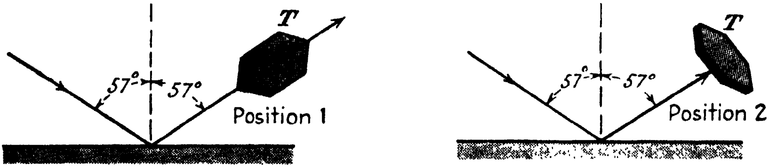


FIG. 33.20. Polarization of light by reflection.

We can now derive the following interesting relationship, which is known as **Brewster's law:** From Fig. 33.21 we see

$$p + 90° + r = 180° \qquad \therefore r = 90° - p \tag{1}$$

According to the law of refraction: $\dfrac{\sin p}{\sin r} = n$ (2)

Substituting equation (1) in (2), we get

$$\frac{\sin p}{\sin r} = \frac{\sin p}{\sin (90 - p)} = \frac{\sin p}{\cos p} = \boxed{\tan p = n} \tag{33.5}$$

Equation (33.5) suggests an interesting possibility. It enables us to determine the index of refraction n of an opaque body. We have merely to examine the degree of polarization of the reflected rays for various angles of incidence in order to find the angle p for which the reflected ray is totally polarized, *i.e.*, totally extinguishable by an "analyzer."

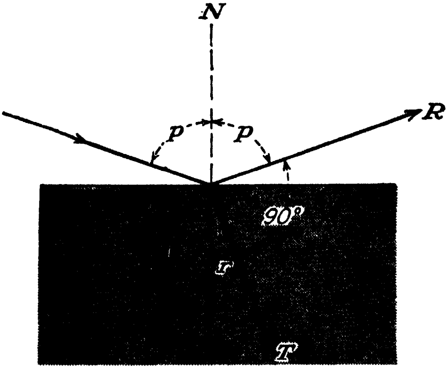


Fig. 33.21.

What does the phenomenon of polarization reveal about the nature of light? If we consider light to consist of a swarm of corpuscles, we can explain these observations by a mechanical hypothesis, by imagining the corpuscles to be, say, disk-shaped. The tourmaline crystal could be imagined to act as a gridlike sieve. Figure 33.22 shows a "beam of disks" moving toward the polarizer P. The disks are oriented in various ways with respect to the plane of the paper. Only those disks manage to slip through the meshwork of the polarizing grid P which are oriented very nearly parallel to the grid wires of P and to the direction of motion. The transmitted swarm can pass freely through the analyzing grid A_1, which is oriented parallel to the wires of P, but is blocked by the analyzer A_2, which is oriented at right angles to the wires of P ("crossed polarizers").

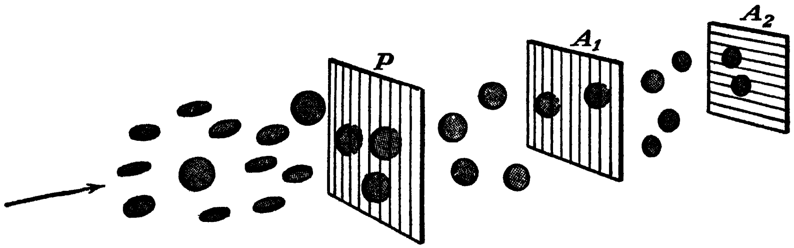
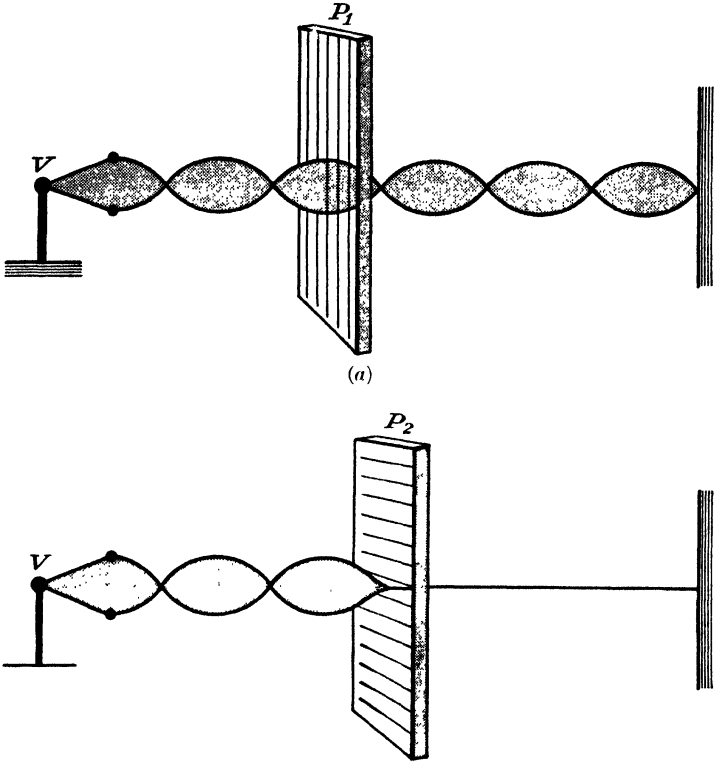


Fig. 33.22.

If, on the other hand, we adhere to the wave theory of light, we can make polarization plausible by the following mechanical analogy. Figure 33.23*a* shows a string, one end of which is maintained in oscillatory motion by the vibrator *V*. If the vibrator moves the string up and down, a standing wave is set up in the string in which all particles vibrate in the vertical plane. The wave is said to be polarized in the vertical plane. A wire grid P_1 through which the string passes does not impede the wave motion if its wires are oriented parallel to the plane of vibration (Fig. 33.23*a*). But when the grid is turned through a right angle, as shown in Fig. 33.23*b*, the wave does not proceed beyond the grid.



(*a*)

(*b*)

FIG. 33.23.

We can imagine rays of unpolarized light to behave like strings which vibrate in different planes. When these strings are passed through a grid, such as P_1 in Fig. 33.23*a*, only the component of the string waves vibrating in the vertical plane will be transmitted, whereas the other (perpendicular) component will be blocked in the manner shown in Fig. 33.23*b*. (Each vibration can be resolved into a vertical and a horizontal component.) This sifting process offers a clear mechanical picture of polarization of waves. *The original beam of light containing waves in which the vibrations occur in all conceivable planes is said to be unpolarized.*

Though this explanation of the polarization of light may sound satisfactory, it nearly led to the downfall of the wave theory of light. The wave theory gained wide acceptance because of the elegance with which it explained the phenomena of interference and diffraction by assuming light to be a wave motion propagated through a *fluid* medium, the "luminiferous ether." Although we never stated explicitly what kind of waves the ether waves were supposed to be, the consideration of the ether as an elastic *fluid* implied the tacit assumption that the light waves were longitudinal ether waves, since only longitudinal waves are conceivable in a fluid. The explanation which we gave for the polarization of light waves suggests, however, that the ether waves are transverse. *Polarization of light cannot be explained on the basis of the wave theory without assuming light to be a transverse wave motion. Thus we are facing a difficult decision. Either we must give up the wave theory of light, or we must give up the idea that the ether is a fluid.* A solid ether would be quite capable of transmitting transverse waves, and an explanation of the polarization of light could be offered by the wave theory.

The assumption of a solid ether looked ridiculous to many outstanding scientists. Laplace and Arago, among others, offered ingenious explanations of polarization on the basis of the corpuscular theory of light, and the fate of the wave theory appeared to be sealed. But a young Frenchman, Augustin Fresnel, refused to accept defeat of the wave theory. He had the courage to adopt the assumption of an *elastic solid ether* which enabled him to regard light as a transverse wave motion in the ether. The brilliant sequence of correct predictions to which Fresnel was led by his assumptions secured him a place among the immortal geniuses of physics and saved the wave theory of light as a useful logical structure based on an unintelligible foundation (namely, on the assumption of an elastic solid and yet penetrable ether).

The objections which were raised against the assumption of a solid ether are quite understandable. How could the planets ever move through it without the slightest sign of retardation during the period of recorded astronomical observation? This question received an ingenious answer from Lodge. He argued that there are certain substances which can behave as a fluid or as a solid, according to circumstances. For instance, pitch behaves like a fluid when subjected to a slow deformation—it will flow through a funnel in the course of months—but when we attempt to bend it quickly by delivering a hammer blow to it, it shatters like a brittle solid. Now the ether could behave in a similar fashion, acting as a fluid with respect to the relatively slowly moving celestial bodies and as a solid with respect to the very rapid vibrations which constitute light.

But these arguments did not remove all the difficulties. Poisson showed that a disturbance inside a solid giving rise to a wave motion should create two simultaneous waves, a transverse wave and a longitudinal wave, the two having, in general, unequal speeds. The longitudinal waves would be recognizable by being nonpolarizable. No radiation was ever observed, however, which could be interpreted as being the anticipated longitudinal ether wave. To account for the fact that longitudinal ether waves have not been observed, Cauchy suggested that their velocity of propagation may be zero, which means that they could not leave the source. But mechanical computations showed that a stable elastic medium in which the velocity of propagation of longitudinal waves would be zero is quite inconceivable.

Solution of the Dilemma through Maxwell's Theory of Light

Maxwell's theory of light appeared as a salvation in this increasingly confusing situation. It offered quite naturally an explanation of the polarization of light without the necessity of accepting the mechanical monstrosity of the "permeable elastic solid ether."

Figures 33.7 and 33.11 illustrate the asymmetric nature of an electromagnetic wave. Such a wave train possesses the kind of "sidedness" we have to assume in order to explain observations such as the extinction of light by tourmaline, illustrated in Fig. 33.19. Thus the existence of polarized light follows quite naturally from Maxwell's theory without the necessity of making highly artificial assumptions. It has been actually shown that light can be polarized by a fine wire grid (compare Fig. 33.13) in a similar fashion as it is polarized by a tourmaline plate (see Chap. 34). This can be considered as very strong evidence in favor of the assumption of identity of electromagnetic waves and light.

Additional successful predictions convinced physicists that Maxwell's theory of light was a step in the right direction. Thus, for instance, equation (33.4) implies a relation between the dielectric constant and the index of refraction of a material; *i.e.*, it links an optical property to a purely electrical one! If we limit ourselves to dielectrics, we can assume with an excellent degree of approximation $\mu = 1$. Then we obtain

$$v = \frac{c}{\sqrt{K}}$$

If we now remember that the index of refraction n of a substance is defined as the velocity ratio $n = c/v$, we can write

$$\boxed{\frac{c}{v} = n = \sqrt{K}} \tag{33.6}$$

This is Maxwell's famous relationship between n and K, which can be tested experimentally. The following table gives a comparison between the experimental values of n and of $\sqrt{K}$ for various substances.

Substance	n	$\sqrt{K}$
Air	1.000294	1.000295
Hydrogen	1.000138	1.000132
Carbon dioxide	1.000449	1.000473
Carbon monoxide	1.000346	1.000345
Sulfur	2.166	2.185
Carbon disulfide	1.611	1.615
Toluene	1.491	1.538
Flint glass	1.6	2.6
Water	1.33	8.9

In most of the cases selected, the agreement is very close indeed. In the cases of flint glass and water, however, there seems to be an excessive discrepancy. We must consider, however, that the index of refraction depends on the wave length of light. In glass, for instance, it is larger for blue light than for red light. It is subject to an even greater variation if we compare n for visible light and for Hertzian waves of long wave length. Experiments show that in the cases of water, flint glass, and other substances showing a similar discrepancy between n and $\sqrt{K}$, the two values become equal when determined for the same frequency.† (We find listed in the above table the *optical* index of refraction of sulfur. Sulfur is an opaque substance. How is it possible to determine its optical refractive index?)

We shall now attempt to answer the following questions on the basis of Maxwell's theory of light: (1) How can we picture the processes of emission and absorption of light? (2) In what way does a polarized beam of light differ from an unpolarized beam? (3) How can we determine the direction of the electric and of the magnetic vector in a polarized beam of light?

---

† The reason for the breakdown of Maxwell's relation $n = \sqrt{K}$ for some substances in the region of visible light is the existence of atomic and molecular resonance effects which were not considered by Maxwell, who treated a dielectric as a continuum.

1. We have discussed the radiation from a wire antenna in which electrons oscillate to and fro (Fig. 33.7). The emanating waves are polarized, the magnetic vector being perpendicular to the antenna and the electric vector being at right angles to the magnetic vector. For future discussions, we should like to mention that *the intensity of radiation emitted by the antenna varies with direction.* Most of the radiation emanates at right angles to the antenna wire; none is emitted in the direction of the wire. The directional variation of the radiation intensity is illustrated in Fig. 33.24. The vectors A, B, C, D, and E illustrate how the diagram enables one to determine the intensity of radiation emitted in the direction of these vectors. The length of the vector indicates the radiation intensity.

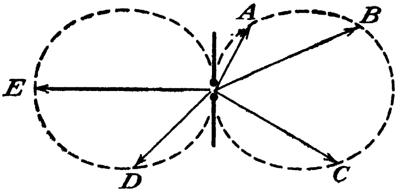


FIG. 33.24. Directional intensity distribution of antenna radiation.

How can we transfer the picture of the radiation process described for an antenna to the emission of light by an individual atom? We know that electromagnetic radiation is emitted whenever there is a change in the antenna current. If the current is due to the motion of a fixed number of electrons, this current will increase when the speed of the electrons increases and it will diminish when the electrons slow down. If the group of moving electrons is reduced to one single particle, then the acceleration and deceleration of this electron constitute the only means of changing the current. Thus we reach the conclusion that an antenna containing a single conduction electron will emit radiation when this electron is accelerated. Since oscillatory motion is an accelerated motion, the antenna will continuously emit electromagnetic waves when the electron performs a simple harmonic motion.

The electrons do not necessarily have to *oscillate* in a wire antenna in order to produce an electromagnetic wave. An electromagnetic "pulse wave" will be produced whenever there is a change in current in a wire. It is easy to visualize how this gives rise to an electromagnetic disturbance which is propagated through space. At a given value of the current there is a definite magnetic field intensity at every point in space. When the current is increased, there is a proportional increase in the magnetic field strength at every point. The "magnetic message of the increase of the current," *i.e.*, the increase in the magnetic field strength, is not established instantaneously in the infinite space but, rather, is propagated with the speed of light in the form of an electromagnetic wave issuing from the antenna.

Let us apply what we have said about emission of electromagnetic waves from an antenna to the emission of an electromagnetic pulse wave from a single accelerated electron. The relation between the emitted wave and the direction of the acceleration vector is similar to the relation between the emitted wave and the antenna in the case of a dipole oscillator. The magnetic lines of force of the emitted wave form a system of rings perpendicular to the acceleration vector. The electric field of the wave is perpendicular to its magnetic field, and the maximum intensity of radiation is at right angles to the acceleration, whereas the intensity of the electromagnetic radiation in the direction of the acceleration is zero. (Compare with the similar radiation diagram for a dipole antenna Fig. 33.24. The direction of acceleration is to be imagined to be parallel to the antenna.)

An oscillating electron can be considered as an idealization of an alternating current flowing in a dipole antenna. It should emit an electromagnetic wave. This suggests the following picture of the process of emission of light. *We imagine electrons to be elastically bound to an equilibrium position inside an atom. These electrons constitute harmonic oscillators whose vibratory motion will lead to the emission of an electromagnetic wave which could be visible light, infrared or ultraviolet radiation according to the frequency of vibration. The line along which the electron vibrates corresponds to the direction of the antenna. No radiation is emitted in that direction. The maximum amount of radiation is emitted at right angles to the direction of vibration.*

When the electromagnetic wave strikes a substance which contains electrons, the alternating electric field of the wave will force the electrons to execute vibrations. These vibrating electrons will now become miniature antennas emitting electromagnetic radiation. *They will reradiate electromagnetic waves.* This process describes **scattering of light.**

The *free* electrons agitated when an impinging electromagnetic wave hits a metal can be considered as an alternating current of a high frequency. If the electric resistivity is high, a large proportion of the impinging energy will be converted into heat (Joule's heat). Thus we find that a metal reflects light the better the lower its electric resistivity. For this reason, silver, which is one of the best conductors of electricity, is also one of the best reflectors of light.

2. The wave emitted by an individual vibrating electron should be polarized in a fashion similar to the wave emitted by the antenna of Fig. 33.7. If we imagine the atoms of a source of light to be oriented in all conceivable directions, we shall expect the wavelets emitted by electrons of different atoms to be polarized in different planes. Thus an observer receiving light emitted by a polyatomic source in a given direction will detect waves with all conceivable orientations of the electric

vector. To polarize this beam we should have to find means to eliminate all waves except those whose electric vector is oriented in the desired direction. This can be achieved by reflecting the wave by a dielectric surface under a proper angle, as is demonstrated by the following experiment. A dipole oscillator A is used as a source of electromagnetic waves (Fig. 33.25a) and is oriented parallel to the plane of a sulfur mirror.

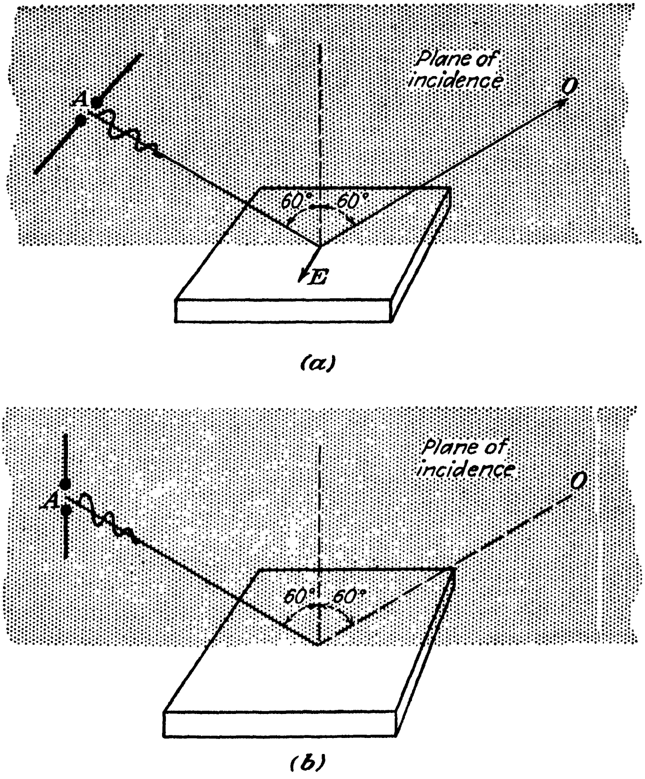


Fig. 33.25a. Reflection of electromagnetic waves. The antenna is parallel to the plane of the mirror. b. Absence of reflected wave when the antenna is parallel to the plane of incidence.

The angle of incidence is 60° (polarizing angle for sulfur). A detector at O registers a strong reflected wave whose electric vector is perpendicular to the plane of incidence, *i.e.*, parallel to the antenna. If we now rotate the antenna A from the position it has in Fig. 33.25a toward the orientation shown in Fig. 33.25b, the detector at O still detects a wave whose electric vector is perpendicular to the plane of incidence, but this wave becomes progressively weaker until its intensity drops to zero when A has assumed the position shown in Fig. 33.25*b*. If we had a mixture of waves of all possible different planes of polarization present simultaneously in the incident beam, the wave reflected toward O could be determined by the following processes: Each incident polarized wave can be resolved into two "component waves": one whose electric vector $E^{\parallel}$ lies in the plane of incidence (Fig. 33.26) and a second wave whose electric vector $E^{\perp}$ is perpendicular to the plane of incidence. The latter of the two component waves is reflected, whereas the former is not. Thus the process of reflection has an effect resembling a sifting action, in which waves of a preferred plane of polarization are transmitted.

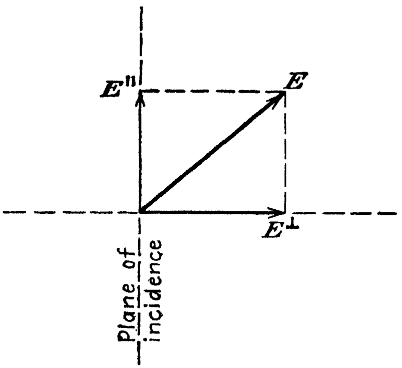


FIG. 33.26. Resolution of the electric light vector E into two components $E^{\parallel}$ and $E^{\perp}$.

3. This experiment also answers our last question. How can we determine the direction of the electric vector in a polarized beam of light? Assume the antenna A of Fig. 33.25*a* to be replaced by a source of polarized light. Imagine then the mirror rotated about the incident beam as axis (the angle of incidence remaining constant and equal to Brewster's "angle of polarization"). In a certain position of the mirror the reflected beam will vanish. We have then the same condition as with the Hertzian waves in Fig. 33.25*b*. Now, if light waves are of the same nature as Hertzian waves, we must assume the electric vector of the polarized incident beam to be parallel to the plane of incidence for the mirror position in which the reflected beam vanishes.

APPENDIX

Radio Communication

The principle of radio broadcasting is based on the emission of high-frequency Hertzian waves by an antenna in which high-frequency alternating currents are maintained. The high-frequency alternating current is generated by an oscillatory LC circuit such as shown in Fig. 33.1. The role of generator is taken over by an amplifier such as the triode circuit shown in Fig. 33.27. This circuit is unstable in the sense that a slight electrical disturbance which gives rise to a weak alternating current in the oscillatory circuit consisting of coil P and condenser C will be amplified to a powerful alternating current of constant amplitude. This is accomplished by electromagnetic induction of an alternating current in the coil S, which acts as the secondary winding of a transformer consisting of coils P and S. The oscillatory variation of the potential difference across the coil S causes the grid voltage V_g to vary periodically. This leads to a periodic variation of the strength of the unidirectional plate current which passes through the triode and is maintained by the battery *B*. The current pulses which pass through the coil *P* are in phase with the alternating current in coil *P* and intensify it. Thus, instead of a damped oscillation as shown in Fig. 33.3, there is an undamped oscillatory current flowing through coil *P*, the energy losses in the coil being replenished by the power delivered by the battery *B*.

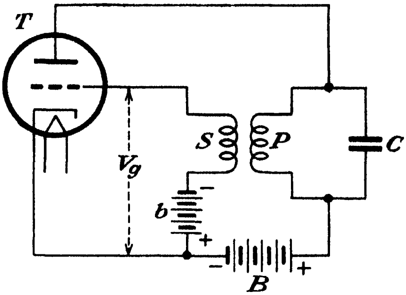


FIG. 33.27. An electron tube oscillator.

The high-frequency oscillations generated in the circuit shown in Fig. 33.27 are used to induce high-frequency currents in an antenna from which a continuous wave train of constant amplitude is emitted (*carrier wave*) (Fig. 33.28*a*). Its amplitude is then *modulated* by a sound wave (Fig. 33.28*b*) which is received by a microphone whose alternating voltage output is applied to the grid of an amplifying tube which amplifies the amplitude of the carrier oscillations before they are applied to the antenna. This varies the amplification of the triode so that the amplitude of the amplified high-frequency alternating current varies as shown in Fig. 33.28*c*. The "envelope" of the high-frequency alternating current output has the shape of the modulating sound wave. Figure 33.28*c* also represents the amplitude distribution in the wave train which travels from the transmitter to the radio receiver.

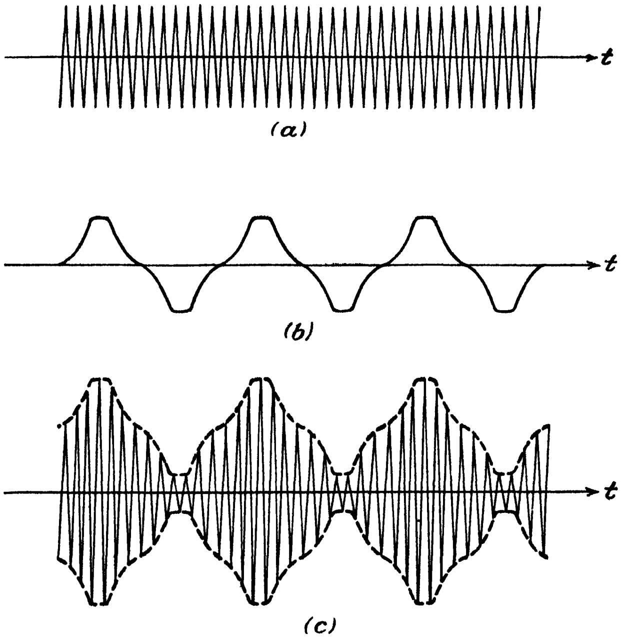


FIG. 33.28. Modulation of a carrier wave. *a*. Carrier wave. *b*. Modulating wave. *c*. Modulated carrier wave.

In the receiving set, the high-frequency a-c signal picked up by the receiving antenna is *rectified* by means of a "diode" which allows the current to pass in one direction only. Figure 33.29 shows how the modulated wave is applied to the input of the receiver. The resistor R is so chosen that the rectified voltage which is applied to the terminals of the condenser C has the shape of the upper half of the wave shown in Fig. 33.28*c*. (The lower half is suppressed because the current can flow only in one direction through the diode T.) This charges the condenser C, which discharges constantly through the resistance R. This resistor is chosen so high that the condenser voltage does not drop to zero between two consecutive pulses of the carrier wave but varies roughly as the *envelope* of the carrier wave. As a result, the current flowing through R has the shape of the original sound wave which was used to modulate the carrier wave. Thus when a telephone (or loud-speaker) is attached to the output terminals O_1 and O_2, one hears the sound which is applied to the microphone at the broadcasting station.

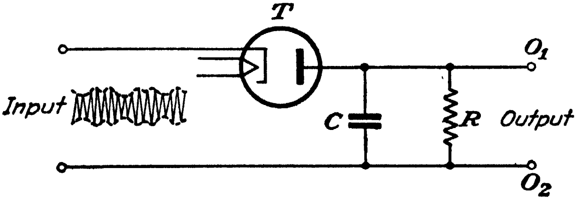


FIG. 33.29. Detector circuit.

QUESTIONS AND PROBLEMS

1. Design a circuit for the generation of electromagnetic oscillations of frequency $f = 10^7$ cycles/sec. Is there more than one set of possible values of L and C that can be chosen?

2. In what way will two oscillatory circuits differ if both have the same values of L and C but the inductance coil of the first circuit has a high ohmic resistance?

3. Can you suggest a more sensitive detector of electromagnetic waves than the spark-gap resonator B_1B_2 of Fig. 33.8?

4. If it were possible to build an oscillatory circuit of zero ohmic resistance, would the electromagnetic oscillations in it persist forever? Why or why not?

5. What will happen to the electrons in a straight wire when the wire is passed by an electromagnetic wave (*a*) when the electric vector is parallel to the wire? (*b*) when the magnetic vector is parallel to the wire?

6. Why can't we have purely electric or purely magnetic waves?

7. Do longitudinal components of the electric or magnetic field occur in plane electromagnetic waves?

8. Explain the results of the experiments shown in Figs. 33.12 and 33.13.

9. In what ways is Maxwell's theory of light superior to the older theory?

10. What is the "polarizing angle"?

11. It is desired to test Maxwell's relation $n = \sqrt{K}$ for hard rubber taking n to mean the refractive index for (*a*) light, (*b*) Hertzian waves. How would you determine the value of n for these two cases?

12. Can polarization of light be explained on the basis of the corpuscular theory? Why were physicists reluctant to give up the wave theory in spite of the difficulties in explaining polarization?

13. Name phenomena which point to a relationship between electricity and optics.

14. Discuss the directional distribution of radiation from (*a*) an antenna, (*b*) an oscillating electron, (*c*) an accelerated electron.

15. Electrons in a betatron (Chap. 24) perform an accelerated motion. Draw the trajectory of a single electron in a betatron, and assuming a constant electron speed, indicate in which direction most of the electromagnetic wave energy will be radiated at a given instant.

16. How does the electromagnetic wave theory explain the emission of nonpolarized light?

17. How can we establish the direction of the electric vector in a polarized light beam?

18. What happens to the energy transported by a plane electromagnetic wave when it passes a group of electrons?

19. Compute the polarizing angle for water.

20. Compute the polarizing angle for sulfur.

BIBLIOGRAPHY

Some interesting applications of polarized light to biological research are to be found in the following references:

Schmitt, F. O., Optical Studies of the Molecular Organization of Living Systems, *J. Applied Phys.*, Vol. 9, p. 109, 1938.

Frey Wyssling, "Submiscroscopic Morphology of Protoplasm," Elvier Co., 1949.

For applications of Hertzian waves to medical therapy, consult Glasser's "Medical Physics," Year Book Publishers, Inc.

CHAPTER 34

EXTENSION OF THE ELECTROMAGNETIC SPECTRUM FROM HERTZIAN WAVES TO X RAYS

a. What experiments demonstrate that light is of the same nature as Hertzian waves?

b. What evidence do we have for the assumption that infrared and ultraviolet light are electromagnetic waves?

c. How are X rays generated?

d. What evidence suggests that they are waves?

e. What evidence suggests that they are electromagnetic waves?

f. Why do we use a heavy metal target in an X-ray tube?

g. How are X rays detected?

h. How can X-ray intensities be measured?

i. What is the physical meaning of the absorption coefficient?

j. On what factors does the coefficient of absorption for X rays depend?

k. On what factors does the absorption of X rays by a block of matter depend?

l. How can X rays be polarized?

m. What hypotheses did Laue's experiment confirm?

n. How does Bragg's X-ray spectroscope work?

o. How can X rays be used to detect the presence of microcrystals in matter? Of what possible medical and biological use is this method?

p. Describe the X-ray spectrum of the radiation originating in an X-ray tube target.

q. How is the X-ray spectrum of the target influenced by the tube voltage?

r. What is meant by the term "characteristic X radiation"?

s. What does Moseley's law state?

The Electromagnetic Spectrum

The evidence which was cited in Chap. 33 in support of Maxwell's electromagnetic theory of light consisted in part of relationships such as (1) $c_W = c_L = \frac{\text{emu of charge}}{\text{esu of charge}}$ (where c_W is the speed of *electromagnetic waves* in a vacuum and c_L the speed of *light* in a vacuum), (2) $v = c/\sqrt{\mu K}$ (v = speed of light in a material medium), (3) $n = \sqrt{K}$ (n = index of refraction, K = dielectric constant), which were predicted by Maxwell's theory and verified by experiment. The results of the experiments of Hertz lent additional support to Maxwell's theory. Hertz was able to show that electromagnetic waves behave like optical waves of a very long wave length. Thus they could be reflected by mirrors, refracted by prisms and lenses, and diffracted by gratings in the same fashion as light. In fact, it proved possible to reproduce practically all the typical optical experiments with Hertzian waves.

To strengthen our conviction that light is an electromagnetic wave motion it would be desirable to show, conversely, that some of the typical electromagnetic phenomena, such as generation of waves by dipole oscillators or polarization of waves with wire grids, could be observed with light and the invisible components of the solar spectrum.

Glagoliewa-Arkadiewa actually succeeded in producing infrared waves of $\lambda \simeq 8 \times 10^{-3}$ cm using suspended iron filings as miniature antennas. These waves are shorter than the longest observed waves of heat radiation ($\lambda \simeq 4 \times 10^{-2}$ cm), so that there is a region of overlap in which we are able to generate waves by optical as well as by electromagnetic means.

The relation between electromagnetic waves and optical radiations is brought out quite clearly by the following experiments:

1. In analogy to the experiment shown in Fig. 33.13, it proved possible to polarize infrared and visible light by means of metallic grids. For experiments with infrared radiation, grids made of fine wire spaced 10^{-2} mm apart were used, and in the case of light, grids were prepared by precipitating threadlike metallic deposits on glass. After passing through one grid (polarizer), the radiation was found to be polarized, the electric vector of the transmitted ray pointing at right angles to the wires. If a second grid (analyzer) was placed behind the first one in a similar fashion to the arrangement of the tourmaline crystals in Fig. 33.19, the ray transmitted by the polarizer could be blocked by orienting the wires of the analyzer at right angles to the polarizer wires.

2. The wire grid G (Fig. 33.13) behaves like a massive, well-reflecting mirror toward the electromagnetic waves which impinge upon it coming from A, while practically no waves are transmitted through G. If, however, we turn the wires of G through 90° (so that they point at right angles to the plane of the paper), the reflected wave vanishes and the radiation of A is freely transmitted through G. If light is a similar kind of radiation, we should expect to find a variation in the intensity of reflected light if we try to reflect *polarized* light by a gridlike mirror while the latter is rotated about the perpendicularly incident beam of light. Such gridlike mirrors have been produced by depositing a fine silver layer on glass and cutting this deposit into thin parallel strips by means of the scratching tool of a dividing engine. The experiment confirmed the expected similarity of behavior between light and electromagnetic waves.

3. It was mentioned in Chap. 33 that the effectiveness of a grid as a polarizer and as a reflector depends on the length of the wires. If we choose the wires originally much shorter than $\lambda/2$ and measure the intensity of the reflected wave while the grid and antenna are oriented as shown in Fig. 33.13, we find that the intensity increases as the length of the wires is increased to the value $L = \lambda/2$ (resonance). A further increase of L beyond this value diminishes the intensity of the reflected beam. Another maximum of reflection is found when L reaches the

value λ. Rubens and Nichols produced tiny grids of different wire length by scratching a silver deposit with a dividing engine as described above. They prepared grids of $L_1 = 6.5\mu$, $L_2 = 12.4\mu$, $L_3 = 18\mu$, and $L_4 = 24.4\mu$.[†] Using infrared rays of $\lambda = 24\mu$ they found that the grids L_2 and L_4 acted as considerably better reflectors of polarized beams whose electric vector was parallel to the scratches than the mirrors L_1 and L_3. Since $L_2 \cong \lambda/2$ and $L_4 \cong \lambda$, this can be interpreted as a resonance effect.

Experiments of this kind have confirmed our belief that Hertzian waves and infrared, visible, and ultraviolet light are radiations of the same kind, differing merely in wave length. Their wave nature is demonstrated by interference experiments and their electromagnetic nature by "resonance-reflection" and "polarization" by wire grids and by the fact that they all travel in a vacuum with a speed which is equal to the ratio of the electromagnetic to the electrostatic unit of charge.

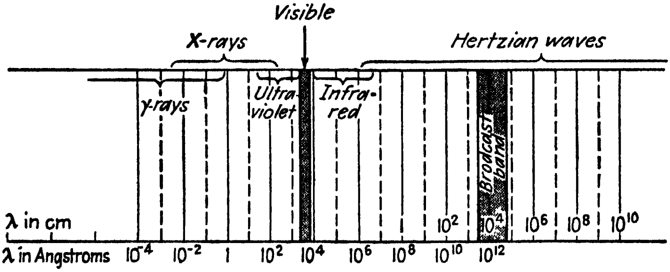


FIG. 34.1. The electromagnetic spectrum.

All these radiations reach our planet from distant corners of the universe. The occurrence of infrared and ultraviolet radiation in the spectra of stars has been mentioned in Chap. 31. Recently also Hertzian ("radio") waves have been discovered which seem to come to us from interstellar space. Their origin and mode of production are still unknown. Figure 34.1 shows a survey of the electromagnetic spectrum.

The applications of visible, infrared, and ultraviolet radiations have been discussed before. The use of Hertzian waves for radio communication is common knowledge. The fact that atmospheric absorption is small and that "short" radio waves are "bent" around the earth [partly owing to reflection from atmospheric ion layers (Heaviside layer)] makes it possible to send radio messages around the earth. The reflection of very short Hertzian waves from obstacles is used in radar to enable us to "see" these obstacles through clouds and fog. A vast number of additional examples could be added to this list of applications. Hertzian waves have proved a powerful tool in acquiring new knowledge and in disseminating information (and sometimes, unfortunately, misinformation).

[†] $1\mu = 10^{-4}$ cm.

The Discovery of X Rays

After seeing how the range of electromagnetic waves has been extended by Hertz's discoveries on the side of the long waves to thousands of meters (long radio waves), one may well wonder whether this extension of the electromagnetic spectrum toward the long waves has not stimulated physicists to search for radiations of very short wave length beyond the ultraviolet end of the solar spectrum. It undoubtedly has, and electromagnetic radiation of frequencies exceeding those of ultraviolet light have, indeed, been discovered. But this discovery did not result from a systematic search for such radiations; rather, it was made accidentally in the course of investigations on passage of electricity through rarefied gases. We described in Chap. 22 how in a low-pressure gas-discharge tube a radiation issues from the cold cathode (cathode rays) and causes the glass to emit greenish light (fluorescence) at the spot where the rays impinge upon it. Figure 34.2 shows a typical cathode-ray tube filled with a rarefied gas. The cathode rays leave the surface of the concave cathode C normally and impinge upon the area A of the glass after crossing at point F. The area A of the glass emits greenish-yellow fluorescent light.

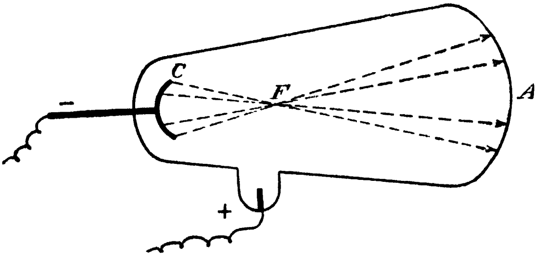


FIG. 34.2. Cathode-ray tube.

The nature of the cathode rays was a much debated subject. Crooks, a leader in cathode-ray research, considered them to be tiny negatively charged particles; Heinrich Hertz, however, believed them to be electrically neutral, since he was unable to deflect them by electrostatic fields. J. J. Thomson succeeded later in finding an inadequacy in Hertz's experimental setup and demonstrated the deflection of cathode rays in electrostatic fields. Lenard was able to design windows of aluminum foil, which he placed opposite the cathode C at point A (Fig. 34.2) and which were sufficiently thin to allow cathode rays to get out of the tube into the atmosperic air, where many of their properties could be studied more conveniently than inside the tube. Among other things, he studied their range in air, penetration of thin foils made of various materials, and their ability to cause various compounds to emit fluorescent light.

Roentgen became interested in the nature of cathode rays and decided to do some research in that field. It was his custom to repeat experiments of his predecessors before beginning explorations of his own. He was interested in particular in comparing the merits of one fluorescent compound of his choice with a fluorescent screen recommended by Lenard for detection of cathode rays in the air. Since the cathode-ray tube emits fluorescent light, Lenard recommended wrapping the tubes in black paper in order to be able to observe without disturbance the fluorescence of specially prepared screens due to cathode rays which escape into the air through an aluminum window. Roentgen followed this instruction with various tubes. But on one occasion he noticed that a fluorescent screen which was not exposed to illumination by the tube lighted up while the tube was operated although there was no light leak in the black paper surrounding the tube. A plausible alternative explanation seemed to be that electrons might have escaped into the air, but this explanation had to be rejected because at the voltages he used the electrons escaping even through the thinnest Lenard windows could not travel more than a few centimeters in air whereas he was able to detect fluorescence more than a meter away from the cathode-ray tube which was wrapped in black paper. Much to his amazement the fluorescence persisted (although less intensely) when sheets of cardboard and even books were interposed between the tube and the screen. The climax of these observations was undoubtedly the moment when he saw the moving shadows of the bones of his hand depicted on the fluorescent screen.

Roentgen realized that he had discovered a new kind of radiation. He made this discovery only a few weeks after beginning to work in the field of cathode rays. Numerous workers, who worked for over a decade in this field, failed to make this discovery because they lacked Roentgen's thoroughness and curiosity. After Roentgen's first publication, several people realized that they had obtained X-ray photographs unintentionally before Roentgen but had not paid any attention to their observations.

The first Roentgen tube was nothing but an ordinary cathode-ray tube, such as shown in Fig. 34.2. The source of the newly discovered penetrating radiation, which Roentgen called **X rays,** was the spot A of the glass upon which the cathode rays were impinging.

X-ray Tubes

It was soon determined that the efficiency of production of X rays could be greatly increased by substituting for the glass spot A a **target** made of a heavy element. Rutherford and Barnes showed in later measurements that the efficiency of X-ray production increases with the atomic number of the target element and with the voltage which accelerates the electrons. The efficiency is, however, quite low even at high voltages. Thus, for instance, at $V = 96{,}000$ volts, only 0.2 per cent of the kinetic energy of the electrons is transformed into X rays, the rest of the energy being converted mostly into heat. In view of the large amount of heat generated, one chooses a good heat conductor as target material, namely, a metal and preferably one that combines the property of a high atomic number with good heat conductivity and a high melting point as well as a low vapor pressure at high temperatures.

Figure 34.3 shows a modern X-ray tube. T is a small disk of the target metal embedded in a copper stem Cu (good heat conductor).

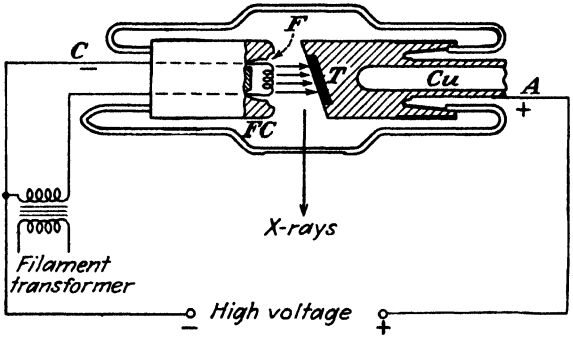


FIG. 34.3. X-ray tube.

The interior of the stem is frequently hollow and is cooled by circulating water or oil. The electrons are derived from a hot filament F, and the tube is highly evacuated. FC is a focusing cup which shapes the electric field inside the tube so as to concentrate the electrons on the target. The accelerating voltage is applied between C and A and ranges in medical applications from about 20,000 volts (skin therapy) to over 10^6 volts (cancer "deep therapy").

Detection of X Rays

Since our sense organs are not sensitive to X rays, we have to detect them indirectly. The most convenient mode of detection is a **fluorescent screen** consisting of uranium glass or of some fluorescent paste deposited on cardboard. These substances are caused to emit visible light by impinging X rays. If such a screen is held in front of an X-ray source, it glows brightly. If a hand is held between the tube and the screen, the outlines of the bones appear on the screen.

If, instead of a screen, we use a **photographic plate** in the above experiment, the shadow of the skeleton appears on the plate after development. Owing to the cumulative effect of long exposures, the photographic plate is a very sensitive (although not a specific) detector of X rays. Besides the decomposition of silver halides, X rays (like light) produce a variety of other chemical reactions which could be used for their detection.

Another very sensitive means of detecting X rays makes use of the **ionization of gases by X rays.** For instance, the leaves of a charged electroscope are made to collapse by X rays traversing the instrument. The discharge is due to the creation of ion pairs in the air which act as carriers of electricity in the electric field between the charged leaves and the oppositely charged housing of the electroscope. Figure 34.4 illustrates a more elaborate arrangement for the measurement of local X-ray intensities as used in medical applications. C is the so-called *ionization chamber* which is exposed to X rays. It consists of a small celluloid "thimble" T (which is made conductive on the inside by a layer of graphite) and of a central graphite electrode R which is insulated from the thimble by an amber plug A. The electrodes R and T of the chamber are connected to a charged electroscope E (the charging circuit has been omitted) through a lead pipe L and a central wire W. Normally, the electroscope deflection remains nearly unchanged after the instrument has been charged. But when X rays fall upon the thimble chamber C, the ion pairs produced in its interior constitute a conductive bridge between the charged leaves and the housing so that the electroscope is discharged. (The lead pipe L limits the ionization effects to the interior of the thimble, since it prevents X rays from entering the space between L and W.) The rate of collapse of the electroscope leaves is used as a measure of X-ray intensity.

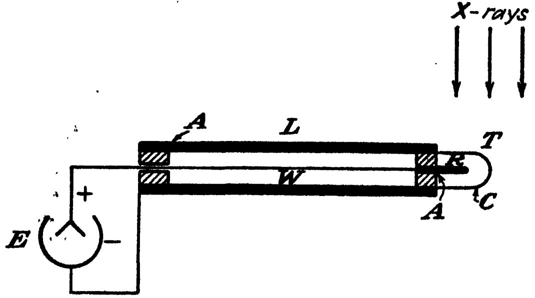


FIG. 34.4. Ionization chamber and electroscope.

Absorption of X Rays

The property of X rays which caused general amazement and attracted world-wide interest is their ability to pass through all known substances, traversing with ease bodies which are opaque to light. The transparency of various substances to X rays can be predicted more simply than transparency toward light. For a given beam of X rays, *the absorption increases very rapidly with the atomic number of the element* (*roughly with the fourth power of the atomic number*). Medical radiography is based on this fact, as a result of which the bones which contain calcium stand out in contrast to the softer tissues, which are composed predominantly of elements of smaller atomic number.†

The penetrating power of X rays varies with the tube voltage (see page 688). *An increase in tube voltage increases the penetrating power of the X rays* (*"hard" X rays*). Radiation produced at a low tube voltage is much more easily absorbed ("soft" X rays). The absorption of X-ray energy is due to various effects which result from the interaction of this radiation with matter, such as *knocking out of electrons* (see Chap. 35 on photoelectric effect and Chap. 37 on Compton effect) and *scattering of X rays* (see below, discussion on polarization of scattered X rays).

The absorption of a radiation by matter is evaluated by comparing the amount of radiation transmitted through a layer of thickness x with the intensity of the unweakened incident radiation (Fig. 34.5). If we consider, for example, a radiation consisting of particles and assume that the number ΔN of particles eliminated from the beam by passing through matter is proportional to the original number N of the particles and to the small thickness Δx of matter traversed, we can write as an expression of this assumption

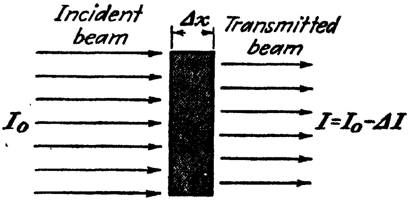


FIG. 34.5.

$$\boxed{\Delta N = -\mu N\,\Delta x} \tag{34.1a}$$

where μ is a proportionality constant, the so-called **coefficient of absorption,** and where the minus sign has been introduced in order to express the fact that ΔN is a decrease rather than an increase in the number of particles (*i.e.*, the slope $\Delta N/\Delta x$ of the curve $N = f(x)$ should be negative: $\Delta N/\Delta x = -\mu N$).

The physical meaning of μ is simple: $\mu = -\dfrac{\Delta N/N}{\Delta x}$. If we set $\Delta x = 1$, we see that $|\mu| = \Delta N/N$; *i.e.*, it is the relative number of particles eliminated from the beam per unit distance.

---

† See discussion of Moseley's law, p. 689, for the definition of atomic number.

The same relationship holds for a radiation which is not considered as a stream of particles. If I is the intensity of the incident beam and $(I - \Delta I)$ the intensity of the beam transmitted through a thin layer of absorbing material which weakens the incident radiation slightly, we can write

$$\boxed{\Delta I = -\mu I \, \Delta x} \tag{34.1b}$$

This equation expresses the fact that the rate of intensity drop $\Delta I/\Delta x$ at any point in the absorber is proportional to the local intensity of radiation I.

Electromagnetic Nature of X Rays

The discovery of X rays initiated immediately speculation on their nature. One of the first hypotheses was the assumption that they were the same type of radiation as visible or ultraviolet light, differing from those radiations merely in wave length. But this idea was abandoned after the failure of the initial attempts to demonstrate their diffraction by slits, refraction by prisms, or polarization by reflection.

Roentgen suggested that X rays might be the long-sought longitudinal vibrations of the ether which Fresnel's theory of light led one to expect in the elastic solid ether. Others suggested that these rays were not ether waves but rather ethereal vortices; Bragg, an outstanding worker in this field, suggested that X rays might consist of a stream of neutral particles. Schuster, however, argued that the original wave hypothesis was not disproved by the unsuccessful attempts to demonstrate diffraction and refraction of X rays. This failure could be explained by the assumption that the wave length of X rays is very small, so that the slits used were too wide to give a clear diffraction pattern, and the absence of noticeable refraction could be due to an index of refraction differing but little from 1. Subsequent research proved Schuster to be right.

In order to prove that X rays are electromagnetic waves, it is necessary to demonstrate (1) their electromagnetic nature and (2) their wave nature. We have accepted as evidence of the electromagnetic nature of light, infrared and ultraviolet rays, the fact that their velocity in a vacuum is $c = \dfrac{\text{emu of charge}}{\text{esu of charge}}$. If X rays are of the same nature as these radiations, we should expect them to have the same speed in a vacuum. Marx designed an ingenious circuit in order to compare the speed of an X-ray signal with the speed of a pulse of Hertzian waves. He found that both travel with the same speed in a vacuum, which supports our assumption regarding the relationship between light and X rays.

The very mode of generation of X rays betrays their electromagnetic character. Maxwell's theory predicts radiation from accelerated (or decelerated) charges. X rays issue from the target *T* (Fig. 34.3) which stops the fast-moving electrons impinging upon it. The magnetic field, which is associated with the moving electrons, vanishes as soon as they are stopped, thus giving rise to an electromagnetic disturbance which is propagated through space. During the period of deceleration the electron emits an electromagnetic pulse wave whose intensity is greatest at right angles to the direction of the acceleration vector. This means that in Fig. 34.3 the X rays are emitted predominantly at right angles to the direction of incidence of the electrons.

A decelerated electron can thus be considered as a miniature antenna. In the example illustrated in Fig. 34.3 the magnetic vector of the emitted wave is perpendicular to the plane of the paper (circular magnetic lines of force surround the electron current flowing from *F* to *T*) whereas the electric vector has the direction of the acceleration vector, *i.e.*, is parallel to the tube axis. Thus, according to the electromagnetic theory, the X rays issuing from the tube should be polarized. This prediction is confirmed by experiment, but the polarization turns out to be incomplete. This result can be explained by assuming that the electrons are deflected by atoms of the target in various directions in the process of deceleration and suffer so many collisions before they are "stopped" that the direction of the tube axis is not the only direction of deceleration.

Polarization of X rays in a manner which made the assumption of their electromagnetic nature very probable was demonstrated by Barkla as follows. He made use of the fact that X rays are scattered diffusely by bodies upon which they impinge and of the prediction of the electromagnetic theory that the scattered radiation should be polarized. His polarizer, as well as the analyzer, consisted merely of a cone† of graphite which scattered the impinging X rays. The polarization of X rays (and, in fact, of any electromagnetic radiation) by scattering is predicted by the electromagnetic theory on the basis of the following consideration: In Fig. 34.6 *S* is a source of unpolarized electromagnetic waves. We indicate the fact that the wave traveling along the *Z* axis is not polarized by drawing a "star" *St* of "double vectors" parallel to the *X*-*Y* plane. This star symbolizes that the unpolarized wave can be imagined to consist of separate polarized wave trains whose electric vectors have the direction of the "double vectors" of the star. *We shall single out one of these polarized wave trains as a typical example for subsequent considerations.* The double vector *E* symbolizes the alternating electric-field vector of this particular polarized wave train.

---

† A conical shape is not essential.

Let us consider what should happen when this wave train passes an assembly of electrons which are associated with a group of atoms. Consider a single electron at point O. As the polarized wave passes it, the electron is "shaken" to and fro along the line EE. This oscillating electron can be considered as a miniature antenna which radiates electromagnetic waves into space. This process of reradiation of electromagnetic wave energy is referred to as **scattering.**

Any displacement of the electron which oscillates along the line EE in the X-Y plane can be resolved into two components, a horizontal one along the line HH and a vertical one along the line VV. If we wish, we can consider the electron as executing simultaneously two vibrations, one along the X axis and the other along the Y axis. It should be equivalent, therefore, to two mutually perpendicular dipole antennas, one parallel to YY and the other one to XX.

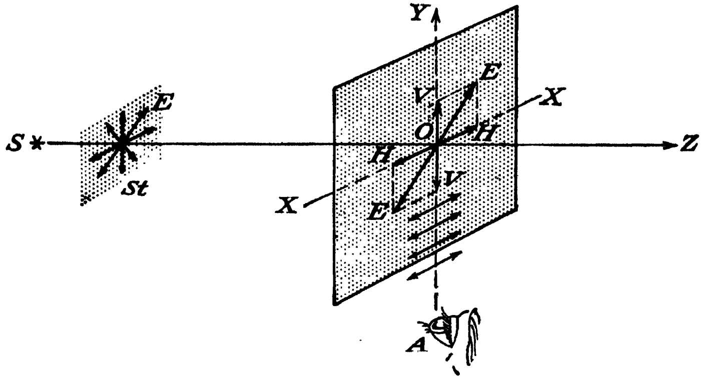


Fig. 34.6. Scattering of electromagnetic radiation.

A glance at Fig. 33.24 reminds us that dipole antennas do not emit radiation in the direction of their own axis, the maximum intensity of radiation being perpendicular to the antenna wire. Hence, we conclude that, of the two "component antennas" (HH and VV), *the vertical one does not emit any radiation in the direction of the observer at A whereas the horizontal one does so with maximum intensity.* The radiation issuing from an oscillating electron vibrating along the line HH is polarized so that the electric vector is parallel to the X axis. *Hence, the observer at A will receive a polarized wave whose electric vector is parallel to HH.*

Instead of the double vector E in the vector star St symbolizing the unpolarized wave, we could have selected one of the other double vectors corresponding to polarization in a different plane. The new double vector would have a different inclination toward the X axis, but the same process of resolution of vibrations into vertical and horizontal components can be applied to this case.

After this introduction we shall be able to understand Barkla's experiment shown schematically in Fig. 34.7. We shall consider the X-ray tube T as a source of nonpolarized X rays which pass through an opening in the lead screen S. The double vector EE represents the electric vector of a typical polarized wave train selected for consideration from the bundle of nonpolarized waves. C_1 and C_2 are two graphite cones which act as scatterers of X rays. When our wave train reaches the cone C_1, its electrons are made to oscillate in the direction of EE.† We resolve this oscillation as above into two components H and V. We now consider the cone C_2 as an "observer." It will "see" only the wave which is emitted from C_1 *owing to the vertical component of the electronic vibrations*, since horizontal vibrations (HH) occur along the line of sight (C_1C_2) and do not lead to emission of electromagnetic waves in that direction (compare Fig. 34.6). Thus the wave scattered from C_1 in the direction of C_2 is polarized, its electric vector being represented by V_1.

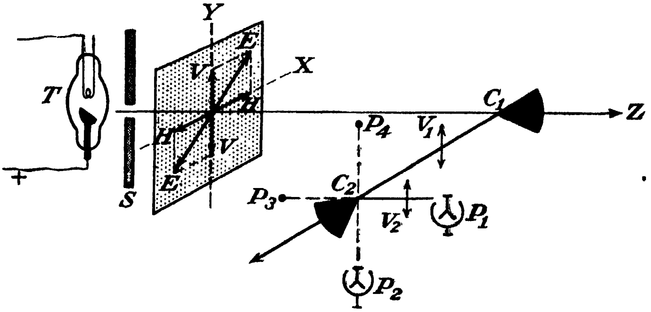


FIG. 34.7. Barkla's demonstration of polarization of X rays.

When this wave (polarized in the vertical plane) reaches the cone C_2, its electrons are forced to vibrate along a vertical line. Hence, an observer at P_1 will detect a scattered wave, but an observer at P_2 will not, since he is located along the line of vibration of the electrons which do not radiate in that direction. Similarly, an observer at P_3 will detect scattered rays whereas an observer at P_4 will not.

Thus the electromagnetic theory of scattering predicts that cone C_1 (polarizer) will become a source of polarized X rays and that the intensity of X radiation scattered from cone C_2 (analyzer) will be different in different directions. Barkla moved an electroscope in a circle about the line C_1C_2 as an axis and found the highest rate of discharge in the positions P_1 and P_3 and a negligible rate of discharge in the positions P_2 and P_4. His success in demonstrating polarization of X rays can be considered a strong support for the assumption that they represent an extension of the electromagnetic spectrum. Similar experiments on polarization by scattering can be performed with light.

---

† Imagine the X-Y plane containing the EE vector transferred to the tip of the cone C_1.

The Wave Nature of X Rays

Having presented evidence of the electromagnetic nature of X rays, we shall now consider experiments demonstrating their wave nature. Experience with all kinds of waves indicates that the velocity of propagation is influenced by the medium in which the wave travels. As a result we are justified, in general, in expecting to observe refraction when waves of any kind pass obliquely from one medium into another. Failure to detect refraction of X rays would leave an undesirable gap in a wave theory of X rays.

Refraction of X rays by a prism was actually demonstrated relatively recently by Siegbahn and his coworkers. The numerical value of the refractive index $n = 0.999995$ which they found with X rays of a particular quality using a glass prism differs so little from unity that it is not surprising that the initial attempts to observe X-ray refraction proved unsuccessful. Another feature of this value of n which will undoubtedly surprise the student is that $n < 1$, so that the wave velocity of X rays must be assumed to be higher in glass than in a vacuum. (The X-ray beam is refracted *toward* the "refracting edge" of the prism rather than away from it. The latter is what we observe with a beam of light.)

But the demonstration of refraction is far from being a convincing proof of the wave nature of X rays. This effect could also be explained on the basis of a corpuscular theory. In our study of sound and of light, we have accepted as conclusive proof of the wave nature of a process our ability to demonstrate standing-wave patterns resulting from interference of waves. We shall search for similar phenomena in the case of X rays to demonstrate their wave nature.

In Chap. 29 (Fig. 29.9), we have shown how an interference pattern is formed when light is allowed to pass through a thin slit. The failure of duplicating this experiment with X rays did not necessarily prove that they are not waves. Waves of exceedingly small λ (as compared with the wave length of light) would give a scarcely perceptible diffraction pattern if slits of the same width as those used with light were to be used with X rays. Not knowing what the appropriate slit width would be for X rays, Haga and Wind (1899) decided to use a wedge-shaped slit (Fig. 34.8), which could be considered as a superposition of slits of different width (sections between the dotted lines) down to zero width. When they sent X rays through a wedge-shaped slit in a screen *S*, they were able to record the outline of the wedge on a photographic plate placed behind the screen *S*. The contour of the triangle on the photographic plate was not sharp at the bottom near the point. The tip of the triangle appeared somewhat widened. This is what we should expect to happen if X rays are waves of a very small λ. In that case only the narrower bottom portion of the wedge would give an appreciable diffraction pattern. But the characteristic bright and dark fringes of diffraction were missing. This did not render the experiment inconclusive; the absence of fringes could be simply explained by assuming that the X rays used were not of a uniform wave length and that the fringes corresponding to different wave lengths were overlapping and causing a smeary pattern. The magnitude of the observed effect allowed an estimate of the order of magnitude of the wave length of X rays. It was concluded that, if X rays were actually waves, their wave length should be about 1/1,000 of the wave length of light. This means that, in order to demonstrate diffraction phenomena by slits and gratings as clearly as we have been able to do in optics, we should use slits and gratings about a thousand times as fine as those we used with light. In other words, we should have to substitute for the optical gratings having thousands of lines per inch gratings of millions of lines to an inch! It is obviously hopeless to rule such a grating by machine.†

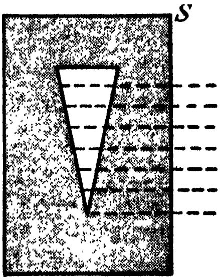


Fig. 34.8.

Laue conceived the idea of utilizing the regular spacing of atoms in crystals as gratings ruled by nature. By passing X rays through a crystal he was able to demonstrate the wave nature of this radiation most convincingly and to determine its wave length with good accuracy. The detailed theory of **Laue's method of X-ray diffraction** is too involved to be discussed here, but a simplified explanation will suffice to elucidate the basic idea.

In Chap. 29 we discussed the interference pattern of a double slit (Fig. 29.3). A similar pattern is obtained when, instead of using two slits, we use two adjacent fine holes in a paper screen. These holes can be considered to act as two coherent sources of light when illuminated by a common source from the back of the screen. The Huygens' wavelets issuing from these holes form a standing-wave pattern by interference.

---

† Lately it has proved possible to demonstrate X-ray diffraction with optical gratings at grazing incidence of the radiation.

This interference pattern is shown by a pattern of dark and bright fringes on a screen which is illuminated by the double slit or "double hole."

Let us now consider two atoms in the path of a wave of X radiation. Electron oscillations will be excited in each of the atoms so that the two atoms will become coherent sources of secondary spherical X-ray wavelets. These secondary wavelets will interfere with each other and should produce an interference pattern similar to the one obtained with a double hole. A row of regularly spaced atoms in the path of the X-ray wave would act like a row of coherent X-ray emitters. The interference pattern due to the row of atoms should be related to the pattern due to a pair of atoms just as the interference pattern caused by a diffraction grating is related to the pattern due to a double slit. More complicated interference patterns are to be expected with several parallel rows of atoms (Fig. 34.9). We should expect a beam of X rays aiming away from the reader to produce a rather complicated interference pattern on a screen behind the paper.

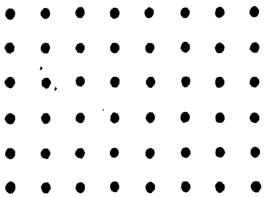


FIG. 34.9.

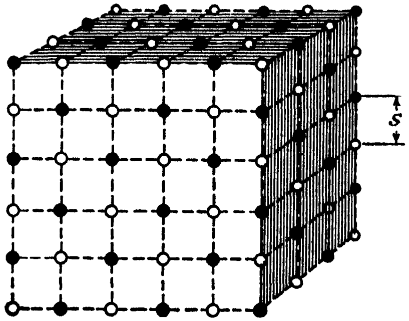


FIG. 34.10. Model of an NaCl crystal.

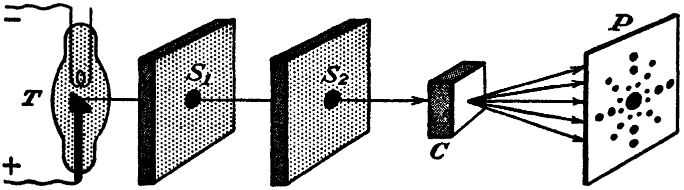


FIG. 34.11. Laue's setup for diffraction of X rays.

If we now imagine many planes studded with atoms (such as the plane shown in Fig. 34.9) placed at regular intervals parallel to each other, we shall obtain a model of a simple crystal (Fig. 34.10). When an X-ray wave passes through such a crystal, the individual atoms of this *space lattice* can be considered as sources of coherent waves. The anticipated interference pattern formed by these wavelets is quite complicated. Laue derived equations which enabled him to compute the wave length of X rays from the observed interference pattern.

The arrangement of the famous experiment, which was performed at Laue's suggestion by Friedrich and Knipping, was very simple (Fig. 34.11). Small holes in the lead screens S_1 and S_2 select a thin pencil of X rays issuing from the tube T. This narrow beam traverses a crystal C and is partly transmitted. The transmitted beam falls upon the photographic plate P and produces a heavy black spot at the center. Owing to diffraction, however, a large number of weaker spots appears on the plate P arranged in a regular pattern, such as shown in Fig. 34.12. Knowing the relative distance between the atoms in the crystal, Laue was able to determine that the X-ray beam he and his coworkers used contained wave lengths in the range of 0.127 to 0.483 A. (Compare it with the wave length of violet light which is 4,000 A.)

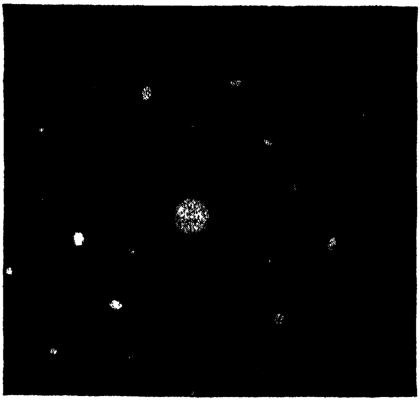


FIG. 34.12. A "Laue photograph" obtained with an NaCl crystal. (*From Compton and Allison, "X-rays in Theory and Experiment," Van Nostrand.*)

This experiment was significant in more than one respect:

1. It demonstrated the wave nature of X rays.
2. It enabled us to determine λ, the wave length of X rays, by measuring the distances between the "Laue spots."
3. It confirmed the idea that crystals consist of regularly spaced atoms.
4. It showed a way to determine interatomic distances in crystals by means of X rays of a known wave length.

Computation of Interatomic Distances

We mentioned that Laue had to determine interatomic distances in a crystal, in order to evaluate λ from the observed diffraction pattern. How can it be done? Consider, for example, the model of a crystal of rock salt NaCl (Fig. 34.10). The sodium atoms are represented by black circles, and the chlorine atoms by white ones. Rock salt crystals are cubic in structure. The smallest cubical rock salt crystal we can imagine is a cube consisting of eight atoms, four of which are Cl and the four others Na† (Fig. 34.13).

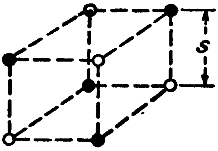


FIG. 34.13.

---

† The atoms of Cl and Na in the rock salt crystal are known to be negatively and positively charged, respectively. The crystal is said to be an *ion lattice*.

We have to exert now a small mental effort to visualize the following fact: Each atom of our elementary crystal (Fig. 34.13) is shared by eight adjacent elementary crystals. This can be seen as follows: Let us fix our attention on the corner of the cube, labeled by the letter A in Fig. 34.14, which shows four adjacent cubes. Point A is shared by the four cubes. If we place four identically arranged cubes upon this group of four cubes, we see that now twice as many (namely, eight) cubes meet at point A. Now imagine A to be an atom in the space lattice of Fig. 34.10 and you will visualize that it is a member of eight adjacent cubes. Only one-eighth of this atom belongs to any elementary cube of which it forms a corner.

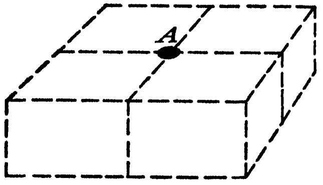


FIG. 34.14.

Now we can answer the simple question: How many NaCl molecules are there in an elementary crystal cube? There are eight atoms in the eight corners of a cube, four Cl and four Na atoms. Only one-eighth of each atom belongs wholly to the cube in question. Hence, we obtain $4 \times \frac{1}{8} = \frac{1}{2}$ Cl atom and $4 \times \frac{1}{8} = \frac{1}{2}$ Na atom as the share of a cube. *Thus we obtain the strange answer to our question: There is one-half of an NaCl molecule per elementary cube in our space lattice.*

We can now compute the distance between a Cl atom and its sodium neighbor (which is equal to the side of the cube of Fig. 34.13) as follows:

The molecular weight M of NaCl is computed from the atomic weights of Na and Cl: $M = 23.0 + 35.5 = 58.5$. The density d of NaCl is known: $d = 2.15$. The number N of molecules in a gram mole of NaCl (58.5 gm) is known (Avogadro's number): $N = 6.02 \times 10^{23}$ molecules/gm mole. The mass of the one-half molecule of NaCl, which is the NaCl content of an elementary crystal cube, must therefore be

$$m = \frac{1}{2}\frac{\text{mass of a gram mol}}{\text{number of molecules per mol}} = \frac{M}{2N} = \left(\frac{58.5}{6.02 \times 10^{23}}\right)\frac{1}{2} \qquad (a)$$

The mass of NaCl contained in an elementary cube can also be expressed in terms of its density d and volume V:

$$m = dV = ds^3 = 2.15s^3 \qquad (b)$$

(where s is the side of the cube = distance between two neighbor atoms). Equating the two expressions (a) and (b) for m we get

$$2.15s^3 = \frac{1}{2}\frac{58.5}{6.02 \times 10^{23}}$$

This yields for the interatomic distance

$$s = \sqrt[3]{\frac{58.5}{2(2.15)6.02 \times 10^{23}}} = \underline{2.81 \times 10^{-8}\ \text{cm}} \qquad (c)$$

Having determined s for one type of crystal, we can calibrate a source of X rays by determining its wave length. Having a source of known λ, we can now determine by diffraction experiments the interatomic spacings of new crystals.

X-ray Spectroscopy

Laue's method accomplishes in a cumbersome way for X rays what a grating spectroscope does elegantly for light. It can be considered to be a method of **X-ray spectroscopy.** Very soon after Laue's discovery, W. L. Bragg invented a much more convenient method of X-ray spectroscopy which is based on the following consideration of the interaction between X rays and crystals: A beam of X rays (ray I of Fig. 34.15), which impinges upon a plane studded with atoms, is partly transmitted (ray T) and partly scattered. Computation shows that, *as a result of the interference between the wavelets scattered by the individual atoms, the intensity of the scattered radiation is greatest in what would have been the direction of the reflected ray* (R) in case the plane of atoms behaved toward the X rays like a mirror. The "reflected" radiation (R) is, however, very weak; most of the X radiation passes without absorption through the atom layer. In the case of an actual crystal we have to imagine millions of planes, similar to the one shown in Fig. 34.15, arranged parallel to each other (compare Fig. 34.10). The part of the radiation which is transmitted through the top layer of atoms is partly reflected by the second layer; still more radiation is reflected from the third layer; etc. This situation is illustrated in Fig. 34.16. The paths Aa, Bb, Cc, . . . drawn from the incident wave front W_1 to the line W_2 are not equal; hence W_2 is, in general, not a wave front, *i.e.*, not a line of equal phase. But if the path difference between two neighboring rays (such as, for instance, Aa and Bb) is an integral multiple of λ, the points a, b, c, d, . . . are points of equal phase and W_2 can be considered as the wave front of a reflected wave. To state it in different words: *For certain values of the angle of incidence α the waves reflected from the different atomic layers according to the optical law of reflection reinforce each other by interference. These values of α are determined by the condition that*

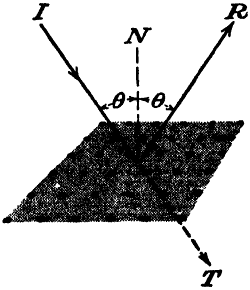


Fig. 34.15.

the path difference 2Δ *between rays reflected from two neighboring atomic layers should be an integral multiple of* λ (**Bragg's condition**).

$$\boxed{2\Delta = 2d \sin \alpha = n\lambda} \tag{34.2}$$

The geometrical significance of d and Δ can be seen from Fig. 34.16. n is an integer.

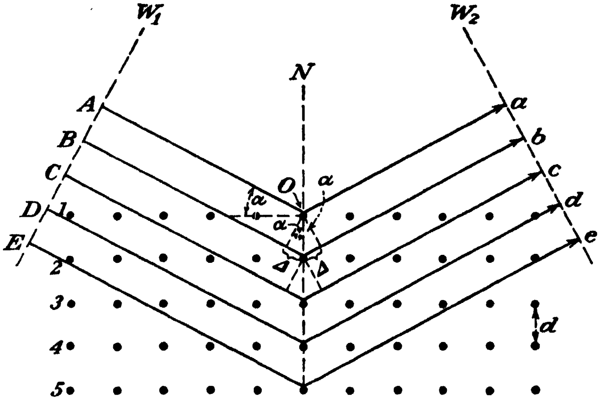


Fig. 34.16.

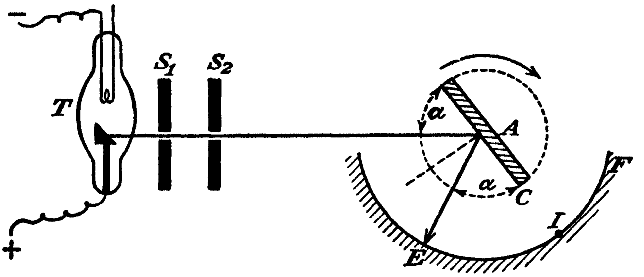


Fig. 34.17. Rotating crystal X-ray spectrograph.

Thus as we vary the angle of incidence, we do not observe a reflected ray at every angle but only at those values of α *which agree with condition* (34.2), *i.e., when* $\sin \alpha = n\lambda/2d$. We obtain the reflection of the first order for $n = 1$, of the second order for $n = 2$, etc. (there is a reflected beam as long as $n\lambda < 2d$).

We are now in a position to understand the principle of Bragg's rotating-crystal method of X-ray spectroscopy. A fine beam of X rays passing through small holes in the screens S_1 and S_2 is allowed to impinge upon a slowly rotating crystal C (Fig. 34.17) (the axis of rotation A is perpendicular to the plane of the paper). F is a photographic film curved along a circular arc. If C were a mirror and the beam were visible light, the end point E of the reflected beam would trace a solid black line, say, from the initial point of incidence I upon the film to the point of incidence E determined by the final crystal position. The reflection of X rays from a crystal differs from the above case by the fact that the reflected beam exists only for certain discrete crystal orientations. *Thus we obtain only a few black dots, instead of a solid black line, in the case of X-ray reflection.* The dots correspond to reflections of the first, second, etc., order (the corresponding values of n being 1, 2, 3, etc.).

In the above example we considered a beam of monochromatic X rays, *i.e.*, rays of a single wave length. Suppose now that our beam contains two different radiations of wave lengths λ_1 and λ_2. If we rotate the crystal only so far as to get only the first-order reflection, we should get two black marks on the film corresponding to reflection of λ_1 and of λ_2. The corresponding values of α are given by $\sin \alpha_1 = \lambda_1/2d$ and $\sin \alpha_2 = \lambda_2/2d$. *We see how the rotating crystal gives spectral lines for X radiations of different wave length similarly to the performance of an optical grating spectroscope.* The openings in the lead shields S_1 and S_2 are usually thin slits, so that the black marks on the film are thin lines.

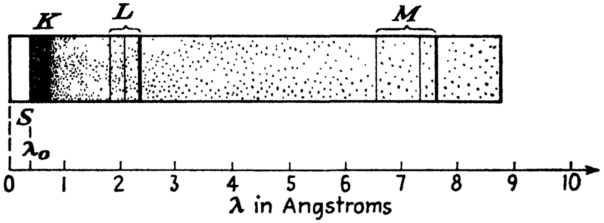


FIG. 34.18. Schematic diagram of the X-ray spectrum of the radiation coming from an X-ray tube target.

Bragg's device enables us to study **X-ray spectra** as we studied spectra of visible light.† Figure 34.18 is a schematic representation of an X-ray spectrum emitted by the target of an X-ray tube. We recognize a sequence of distinct groups of lines labeled as the *K*, *L*, **and** *M* **series.** These lines are recorded on a continuous background of general blackening which extends to a sharp limit S, the so-called **short-wave-length limit.** This observation means that the target emits X rays of certain discrete wave lengths, and, in addition to them, radiation ranging continuously from very long waves to a shortest wave length λ_0.

The short-wave-length limit of the continuous radiation is independent of the target material. It depends in a simple fashion on the tube voltage. This relation is usually expressed in terms of the frequency ν rather than the wave length $\lambda = c/\nu$. The following relation (**law of Duane and Hunt**) was found to hold to a high degree of approximation:

---

† For a theoretical discussion of X-ray spectra, see Chap. 36.

$$\boxed{h\nu_0 = eV} \tag{34.3}$$

or

$$\nu_0 = \frac{e}{h} V \tag{34.3a}$$

since

$$\nu_0 \lambda_0 = c$$

we get

$$\frac{c}{\lambda_0} = \frac{e}{h} V \tag{34.3b}$$

or†

$$\lambda_0 = \frac{hc}{eV} \tag{34.3c}$$

In equation (34.3) V is the accelerating (tube) voltage in electrostatic units, e is the charge of the electron in esu of charge, ν_0 is the high-frequency limit (corresponding to the short-wave-length limit λ_0), and $h = 6.62 \times 10^{-27}$ erg-sec is a universal constant (Planck's constant).‡

The spectral lines of Fig. 34.18 *are referred to as the* **characteristic radiation,** *since their position depends on the nature of the target material and not on the tube voltage.*

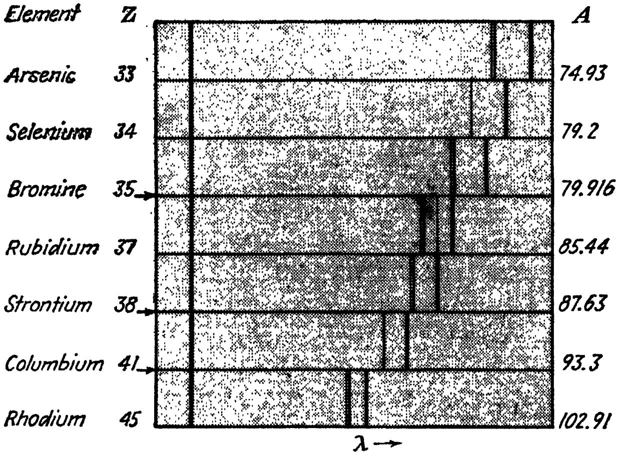


FIG. 34.19. X-ray spectra of the K series of several elements.

Moseley's Law

Moseley found a very simple relationship between the characteristic spectra of elements of which he made a systematic survey. Figure 34.19 shows spectra of the K series of several elements. We see how the wave length of the two lines of the K series diminishes as we progress to heavier elements (the atomic weight is indicated under the letter A in the diagram). This steady change of λ suggested to Moseley the idea of using the X-ray spectra of the elements as a criterion for their arrangement. Thus he arrived at a sequence of elements of progressively decreasing wave lengths of the characteristic K spectrum almost identical with the sequence of elements arranged according to their atomic weight which underlies Mendeléeff's periodic system. But Moseley's sequence differed slightly from the sequence according to atomic weights. For instance, according to the X-ray spectra, argon should precede potassium, although its atomic weight (39.944) is greater than that of potassium (39.10). In arranging the periodic table, chemists were compelled to abandon the atomic weight as a criterion of order in the case of argon and potassium and in a few other cases in order to preserve the correct grouping of elements according to their chemical nature. Moseley's sequence provided a natural justification for that apparently arbitrary decision. Moseley considered his criterion as superior to the original arrangement of the elements according to their atomic weight. He called his sequence of elements the **natural system of elements** and the order number of an element in that sequence the **atomic number Z.** As we shall see later, this number has an important physical significance.

---

† If V is expressed in volts instead of esu, $\lambda_0 = 12{,}340/V$.

‡ For practical computations remember that 1 esu = 300 volts.

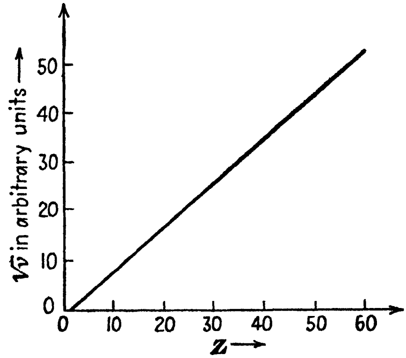


Fig. 34.20. Graphical representation of Moseley's law (a plot of $\sqrt{\nu}$ for the $K\alpha$ line versus Z).

Plotting the square root of the frequency of a definite line of the K series as a function of the atomic number Moseley obtained a straight line (Fig. 34.20) which is described by the equation

$$\boxed{\sqrt{\nu} = a\,(Z - b)} \qquad \textbf{(Moseley's law)} \qquad (34.4)$$

A missing element in the sequence can be easily detected by a break in the linear plot. Even a superficial inspection of the X-ray spectra of Fig. 34.19 indicates the extra large changes in wave length at places where an element is omitted (the gaps are indicated by arrows). Thus, Moseley was able to determine that there are no more than 91 elements preceding uranium in his sequence and to predict elements which were yet to be discovered. The knowledge of the position of the characteristic lines in the X-ray spectrum of an element according to Moseley's law offers a simple means of detecting the new element in a mixture of substances by means of X-ray spectroscopy.

X-ray Structure Analysis

Bragg's method can be used, like Laue's method, for the analysis of crystal structure as well as for X-ray spectroscopy. The necessity of preparing a large crystal for analysis and of rotating it is an inconvenience. Debye and Scherrer worked out a method which allows us to analyze powdered material, containing tiny crystals, with a stationary apparatus. Their arrangement is shown in Fig. 34.21. A narrow beam of X rays traverses the powder sample S. The sample contains thousands of little crystals which are oriented in all conceivable directions. Some of them are oriented with respect to the incident beam, so that the Bragg condition for the angle of incidence is satisfied. Those crystals reflect X rays according to the law of reflection in a definite direction. With a monochromatic beam of X rays we should expect several reflected rays (several values of α) corresponding to different values of n in equation (34.2). The reflected rays R resulting from different orders of interference are shown in Fig. 34.21. They impinge upon a photographic film F bent in form of a cylinder around the sample S. To visualize the pattern recorded on the photographic film F, we can imagine the reflected rays R rotated about the incident ray. This is justified because a crystal reflecting the beam, for instance, toward the reader through the angle α is just as likely to occur as one reflecting it, for instance, upward or sideways through the same angle. The photograph shows a series of rings when the film is cut and straightened out into a plane rectangle (Fig. 34.22). The spacings and relative intensities of these rings reveal the structure of the microscopic crystals present in the powder. Even submicroscopic crystals yield a diffraction pattern. Thus, for instance, it was possible to demonstrate the presence of microcrystals in liquids.

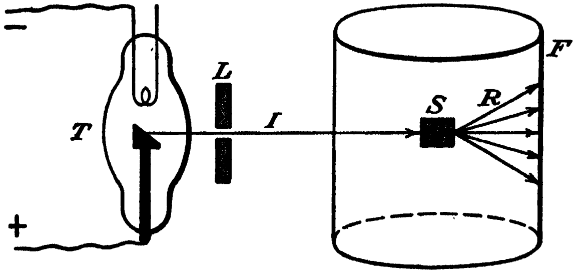


FIG. 34.21. The Debye-Sherrer method of X-ray spectroscopy.

This method of **X-ray structure analysis** promises important biological and medical applications. Diffraction patterns have enabled pathologists to detect in certain cases subtle structural changes in abnormal tissues which could not be recognized otherwise. Interesting results have been also obtained in studies of virus crystals, of tissue development in growing embryos, and of changes in nucleo proteins. This technique may become a powerful supplement to the microscope in morphological analysis.

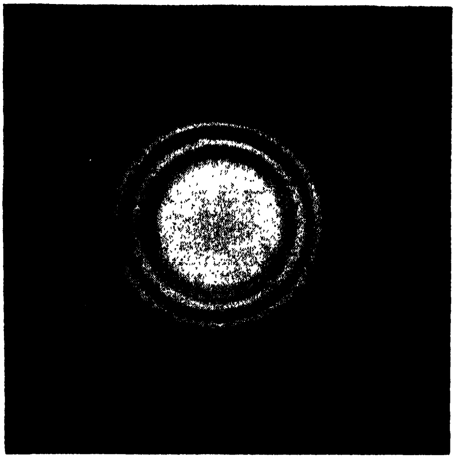


FIG. 34.22. X-ray diffraction photograph obtained with an X-ray beam passing through an aluminum foil containing microcrystals. (*Hull's photograph reproduced from Compton and Allison, "X-rays in Theory and Experiment," Van Nostrand.*)

Additional Examples of Medical Applications of X Rays. In addition to the above-mentioned structure analysis by X rays there are some important biological applications of X rays to be mentioned.

Radiography. The fact that the X-ray coefficient of absorption increases roughly with the fourth power of the atomic number is responsible for the great contrast which we observe between the transmission of X rays by soft tissues and by bones, which are rich in the relatively heavy calcium. Besides revealing bones, the X rays allow us to view various internal organs, such as the heart or stomach, during the performance of their function in health and disease. Frequently substances containing heavy elements (such as "thorotrast") are injected into the blood stream or swallowed in order to delineate more sharply certain organs in which these "radio-opaque" substances accumulate.

In order to obtain sharp shadows, the diagnostic X-ray tube must have as small a target as possible. This presents a technical difficulty, since, when too much energy is delivered in a short time to a small target, it may melt and partly evaporate. Evaporation of the target deposits on the exit window of the tube a layer of tungsten which strongly absorbs the radiation. These difficulties have been solved remarkably well in commercial X-ray tubes.

Radiotherapy. One of the principal therapeutic uses of X rays is in the treatment of cancer. This therapy is based on the fact that, whereas X rays kill both healthy and cancerous tissue, by suitable distribution of the radiation complete destruction of the cancer is achieved occasionally without excessive loss of vital tissues, which may regenerate later on.

X-ray deep therapy makes use of very penetrating radiation produced by applying a high voltage to the tube. Million-volt generators have been frequently used. Betatrons accelerating electrons, which bombard a target, to more than 100 million volts are now available for experiments in Roentgen therapy.

The germicidal action of X rays is utilized in treating various fungus infections of the skin. "Soft" X rays are used in this application. They are generated at low voltage and are absorbed in superficial skin layers.

When an alternating voltage is applied to a high-voltage X-ray tube, soft X rays are produced during part of the time, namely, during the low-voltage period of the cycle. This radiation may give rise to serious skin damage and must be removed lest the administrable dose of hard X rays for cancer treatment be limited. This elimination of soft X-ray components is accomplished by absorption in filters of Al and Cu.

Overexposure and Genetic Effects. Overexposure to X rays may lead to loss of hair, changes in blood composition, ulceration of the skin, sterility, and death. It also causes gene mutations. Offspring of subsequent generations may show congenital abnormalities such as missing fingers, extra arms, etc. This effect is due to modification of genes by X rays. It has been studied extensively by Muller on the Drosophila fly whose eggs were exposed to X rays. The first generation did not necessarily show any effects, but their offspring and later generations did.

QUESTIONS AND PROBLEMS

1. Can X rays be generated by impact of (*a*) ions, (*b*) neutral particles upon a target?

2. In what way is the effectiveness of an X-ray tube affected by evaporation of the target?

3. How could you control the X-ray output of a tube without affecting the penetrating power of the emitted radiation?

4. Is the X-ray output of a tube affected by changing the tube voltage?

5. If we compare the absorbability in copper of X rays generated by the following two setups, what difference shall we find when (*a*) 100,000 volts d-c is applied across the tube? (*b*) an a-c voltage of the effective value of 100,000 volts is applied across the tube?

6. What is the advantage of rectifying the voltage which is to be applied to a therapeutic X-ray tube?

7. Aluminum and copper filters (sheets of Al or Cu) are usually placed between the therapeutic X-ray tube and the patient. What is the purpose of these filters?

8. How is the generation of X rays explained on the basis of the electromagnetic theory?

9. If we double the thickness of a copper plate through which X rays pass, do we necessarily reduce the intensity of the transmitted beam by 50 per cent?

10. Which of the following do you consider the *best* evidence for the electromagnetic nature of X rays?

a. Their speed in a vacuum is equal to that of light.

b. They can be polarized.

c. They can be refracted.

d. They knock electrons out of matter.

11. Light scattered from the blue sky is polarized. Explain why. What is the direction of the electric vector?

12. Explain why X rays emitted by an X-ray tube are partly polarized. Why are they polarized? Why only partly?

13. How can X rays be used to determine interatomic distances in crystals?

14. How many orders of reflection can be observed with a Bragg X-ray spectrograph using a radiation of $\lambda = 2$ A and a crystal of (*a*) $d = 6$ A? (*b*) $d = 1$ A?

15. Illustrate how the existence of an undiscovered element can be determined with the aid of Moseley's law. How could one detect this element with certainty in a sample of matter?

16. How much heat is generated per second at the target of an X-ray tube operated at a tube voltage of 100,000 volts and an electron current of 10 ma?

17. An X-ray crystal spectrometer is calibrated by means of an X-ray beam of known wave length λ_1. It is found that the first-order reflected beam forms an angle of 30° with the crystal surface. It is found that the second-order reflected beam of unknown wave length λ_2 is reflected through 45°. Find the ratio λ_2/λ_1.

18. How fast are the electrons which strike the target of an X-ray tube operated at a tube voltage of (*a*) 50,000 volts? (*b*) 200,000 volts?

19. What minimum tube voltage is required to produce X rays of the wave length of (*a*) 1 A? (*b*) 0.1 A?

20. How large is the energy (in ergs) of an electron which has "fallen" through a p.d. of 1 volt, starting from rest (1 electron volt)?

21. Find the shortest wave length of X rays emitted by a tube operated at a voltage of (*a*) 10,000 volts, (*b*) 100,000 volts, (*c*) 1,000,000 volts.

22. The crystal constant (d of Fig. 34.16) of a Bragg X-ray spectrometer crystal is $d = 2.82$ A. The second-order reflected beam forms an angle of 30° with the crystal surface. Find the wave length of the radiation and the angle of deviation of the third-order reflected beam.

23. Find the angle of first-order deviation of the radiation corresponding to the short-wave-length limit obtained with a tube operated at 100,000 volts and the same crystal as used in Prob. 22.

BIBLIOGRAPHY

On biological applications of X rays:

Glasser, D., "Medical Physics," 1 ed., pp. 88, 495, 512, 663, 730, 1267, 1269, 1288, 1335, 1352, 1401, Year Book Publishers, Inc., 1944.

> In order to reach the truth, it is necessary, once in one's life to put everything in doubt—so far as possible.
>
> *Descartes*

PART D: The Failure of the Classical Mechanical and Electromagnetic Theories and the Revision of the Fundamental Assumptions of Physics

After establishing a theory concerning the propagation of light we proceed to the consideration of the processes of light emission and absorption. It appears that a theory of these processes could be based on an electromechanical model of the atom. Periodic motion of electrons in the atom could explain the emission of light as the emission of an electromagnetic wave by an antenna traversed by an alternating current. It seems that application of the laws of mechanics and of Maxwell's electromagnetic laws to such an atomic model might enable one to predict the characteristic absorption and emission spectra of various elements. But the spectral series of atoms do not seem to be explicable by any familiar oscillatory motion. The emitted frequencies do not form harmonic series. And besides, the periodic motion of the electrons, being an accelerated motion, could not persist according to Maxwell's theory of radiation because of constant radiative energy losses.

Bohr succeeded nevertheless in developing a theory of emission and absorption of radiation based on a visualizable atomic model containing accelerated electrons (Chap. 36). He was forced, however, to give up certain fundamental notions of mechanics and to disregard certain consequences of Maxwell's electrodynamics by imposing certain arbitrary restrictive conditions upon the motion of electrons and the emission of electromagnetic radiation by them. Bohr's restrictive conditions of motion were suggested to him by Planck's and Einstein's ideas on the nature of light, which were based on studies of black-body radiation and of the photoelectric effect (Chap. 35). According to these ideas light is emitted and absorbed discontinuously in form of localized "lumps of energy" (photons). Bohr accounted for the discontinuous emission of light by assuming his atomic model to be stable only for certain discrete energy values of the moving electrons. Bohr's theory proved very successful but only for a limited range of phenomena (Chap. 36).

A more successful theory of atomic behavior finally resulted from a more radical step than that which was taken by Bohr (Chap. 37).

Whereas Bohr made use of Newtonian mechanics and merely supplemented Newton's laws by certain restrictive conditions, the new mechanics renounced completely the validity of Newton's laws of motion for particles in the interior of atoms and replaced them by laws which resemble the equations of optics more than those of mechanics. This new so-called *wave mechanics* is based on an extension of the wave-corpuscle dualism (established for photons in Chap. 35) to material particles. The basic assumptions of the new theory are tantamount to a renunciation of some fundamental postulates of science such as the unrestricted "principle of causality." Single events are assumed to be unpredictable, only statistical predictions being possible. The new atomic theory is more abstract than any of the previous theories, being a mere mathematical scheme of prediction.

In Chaps. 38 and 39 studies of nuclear disintegration offer further illustrations of statistical laws and of the unpredictability of single events. Observation of nuclear radiations and reactions leads to the abandonment of some additional fundamental assumptions of classical physics, such as the notion of the immutability of elements, the assumption of constancy of mass, and the belief that the laws of conservation of mass and conservation of energy are two independent laws of nature.

In Chap. 40 the theory of relativity is presented, which predicted, among other striking new phenomena, the observed interconversion of mass and energy and which led to revolutionary changes in our conception of space and time. This theory is an example of a purely logical theory free of any pictorial content. From two fundamental postulates, using a new definition of simultaneity, Einstein drew logical conclusions which revealed new laws of nature and which compelled us to revise our original notions of space, time, mass, and energy.

CHAPTER 35

EVOLUTION OF THE THEORIES OF LIGHT AND MATTER

a. How could you explain the emission and absorption of discrete frequencies of light on the basis of the classical absorption theory?

b. Would you expect the oscillations of an electron to be affected by placing it in a magnetic field?

c. What is the Zeeman effect?

d. Why can the characteristics of spectral series be considered as evidence against the harmonic oscillator theory of the emission and absorption of light?

e. What role did Balmer's series formula play in the development of atomic physics?

f. How did Thomson propose to visualize an atom?

g. Is the attraction of an electron in Thomson's atomic model toward the center of the atom inversely proportional to the square of the distance from the center?

h. In what way does the interaction of an alpha particle with a Rutherford atom differ from its interaction with a Thomson atom?

i. How did Thomson estimate the number of intraatomic electrons? How large did he find this number to be?

j. What experiments and conclusions led Rutherford to propose his nuclear atomic model?

k. How did one infer the magnitude of the nuclear charge?

l. How can we estimate the upper limit of the nuclear radius?

m. How large is the number of atomic electrons compared with the nuclear charge?

n. What difficulties do we encounter in assuming the electrons to revolve about the nucleus?

o. What assumptions did Planck make (1) about the possible values of the energy content of a harmonic oscillator? (2) about the processes of absorption and emission of energy by a harmonic oscillator?

p. How did Einstein use the quantum theory in interpreting the photoelectric effect?

q. What is a photon? What is its (1) energy? (2) momentum?

r. How is the energy of the emitted photoelectrons affected by (1) the intensity, (2) the wave length of the illuminating light?

s. What is the "work function"?

t. What experiments can be cited in support of the photon hypothesis?

The Classical Ideas on the Mechanism of Light Emission

In Part C we have described the gradual evolution of our ideas on the nature of light. We have seen how the corpuscular theory was supplanted by the wave theory. According to the original wave theory, light was a wave motion of the "luminiferous ether." These waves were supposed to be generated in the ether by the molecular motion in luminous bodies. An increase in temperature of the source meant an increase in the molecular agitation, which was imagined to lead to an increase in the emission of radiation.

The difficulties which the elastic ether theory of light met, particularly in attempts to explain polarization, paved the way for the adoption of Maxwell's electromagnetic theory of radiation. According to this theory light waves are generated by an electromagnetic disturbance rather than by thermal agitation of neutral molecules. A vibrating electric charge ~~would be~~ a conceivable source of electromagnetic waves. Generally, any accelerated motion of electric charges gives rise to electromagnetic radiation. Continuous trains of electromagnetic waves could be gener~~ated by~~ any kind of periodic accelerated motion of an electric charge, such as vibratory or circular motion. This picture of generation of light waves by periodic motion of electric charges is quite consistent with preliminary ideas which we have formed of the structure of matter. We have seen that atoms are electrically neutral, *i.e.*, contain equal amounts of positive and negative electricity. The positive electricity was found to be always bound to the bulk of the mass of the atom, whereas negative electricity could be split off in the form of (the relatively light) electrons. We assumed that negative electricity preserves its corpuscular structure inside the atom, so that we imagine atoms to contain electrons. If the electrons are bound to some equilibrium position inside the atom, a disturbance of this equilibrium would lead to electronic oscillations which would generate light. (It is true that the positive part of the atom may also vibrate as a result of the above-mentioned disturbance, but on account of its large mass the amplitude of this vibration is negligible.)

Zeeman Effect as a Confirmation of the Classical Electron Theory of Light Emission

The classical electron theory of the emission of light looked at first exceedingly promising because of the large number of known phenomena which it explained. The faith in this theory was greatly strengthened by successful prediction of new effects, such as the effect of a magnetic field upon the emission of light by a source. Zeeman was able to show that the spectral lines of sodium become appreciably wider when a yellow sodium flame is placed between the poles of a powerful magnet. Lorentz deduced from a mathematical analysis of the periodic motion of electrons in a magnetic field that the widening of the lines observed by Zeeman should be actually due to a *splitting of the spectral lines.* He predicted that an observer looking in the direction of the magnetic lines of force should see two spectral lines instead of each original line, as observed without the field. If the frequency corresponding to the original line is ν_0, the frequencies of the new lines predicted by Lorentz are $\nu_0 + \Delta\nu$

and $\nu_0 - \Delta\nu$, where $\Delta\nu = (e/m)(H/4\pi c)$ (e/m = specific charge of the electron, c = speed of light, H = magnetic field intensity). Thus measurement of the separation ($\Delta\nu$) between the two new spectral lines should enable us to compute e/m, since c is known and H can be measured. If e/m turns out to have the same value as the specific charge of the electron determined by other methods, we have a good reason to feel confident about this theory. Initial refined observations actually confirmed Lorentz's predictions. In particular, the value of e/m computed from the data agreed with the e/m of electrons.

Lorentz predicted, furthermore, that a spectroscopist studying the radiation emitted at right angles to the magnetic field lines should observe three lines for each original line, one of which would correspond to the original frequency, whereas the two other frequencies would be given by the above expressions. He also foretold that the light giving rise to the new lines would be polarized and specified in what way. These predictions were also confirmed by experiment.

Difficulties of the Classical Electron Theory of Light Emission

Lorentz's brilliant success may lead us to suspect that the electron theory of radiation based on the model of an electronic harmonic oscillator was firmly established. This was, however, not at all the case. The triumph did not last long, and we can take this example as a warning against premature acceptance of a theory even if its initial success may look striking. Subsequent experiments performed with more refined optical instruments and a great variety of light sources revealed that spectral lines may be split into six or more lines for which Lorentz's theory was unable to account even in a qualitative fashion.

The weakness of the harmonic oscillator theory of the process of emission of light appears in the consideration of the basic processes of emission and absorption of light. The emission of a variety of spectral lines of different frequencies could be explained by assuming several electrons in an atom which are bound by different forces. (As we recall from Chap. 25, the natural frequency of an oscillator depends on the force constant k.) But an atom consisting of such an assembly of oscillators of different frequencies should not only emit but also resonate to these frequencies and absorb them. We find, however, that the absorption spectra of gases are as a rule not identical with their emission spectra. The absorption spectrum usually shows fewer lines. This fact is difficult to explain on the basis of the above hypothesis.

An even greater puzzle attends the explanation of the line spacing in spectral series. The picture of an atom as an assembly of several electronic oscillators, designed to explain why many spectral lines are encountered in an atomic spectrum, is quite arbitrary and, in fact, inconsistent with known facts. This can be seen most clearly in the example of the hydrogen spectrum. This spectrum shows many lines, and yet we know that a hydrogen atom can contain only one electron because a hydrogen ion cannot have a charge exceeding the charge of one electron. How can this single electron emit light of so many different frequencies? One might try to explain this as the emission of harmonics. We have seen in Chap. 27 how a string can vibrate in different modes, emitting, besides the fundamental frequency f_0, the harmonics $2f_0$, $3f_0$, etc. But most unfortunately, the frequencies emitted by hydrogen cannot be fitted into a harmonic series. They are not integral multiples of a fundamental frequency, but there is an additive relationship between the different observed frequencies such that, for two frequencies, ν_1 and ν_2, we usually observe a frequency $\nu' = \nu_1 + \nu_2$ and a frequency $\nu'' = \nu_1 - \nu_2$ as well. Thus we reach the conclusion that the mechanism of the emission of light presented above is incorrect.

We shall now set out to look for a correct theory. An acceptable explanation of the emission of light should enable us to predict the frequencies emitted by an atom or at least by the simplest atom, namely, hydrogen. Our search for an acceptable theory will proceed along the following lines:

1. We shall study how spectral lines can be grouped in general into spectral series.
2. We shall then learn how mathematical relations can be established between the members of a spectral series from which the frequencies of the various spectral lines of the series can be computed.
3. Our final step will then be to search for a hypothetical model of the atom which would yield, through application of the laws of Newtonian mechanics and electromagnetic theory—possibly in conjunction with special new assumptions—the same expressions for the frequencies of spectral lines as have been found empirically in item 2.

Spectral Series

The study of the wave lengths of the spectral series played the same role in the development of atomic mechanics as the study of planetary motions played in the development of Newtonian mechanics. Early spectroscopists noted certain regularities in the multitude of lines in atomic emission and absorption spectra. They observed that certain groups of lines could be singled out such that the interval between neighboring lines was diminishing toward zero in a regular fashion, the lines eventually converging toward a common series limit (H_∞ in Fig. 35.1). The identification of the lines which belong to a spectral series is frequently a laborious process. It can be accomplished by examination of individual spectral lines by means of spectroscopes of high resolving power. Such studies show that many of the spectral lines are actually *multiplets*, for instance, groups of two (doublets) or of three (triplets) lines. Hartley observed that the frequency difference between the doublet lines of a recognized series remained the same for all lines of the series. He found more generally that the frequency difference between the individual lines of a multiplet remained constant for all corresponding multiplet lines of a given series. Thus Hartley's criteria made it possible to single out members of a spectral series even in complicated spectra.

The recognition of spectral series is particularly easy in the case of the relatively simple spectrum of hydrogen. Figure 35.1 shows the lines H_α, H_β, H_γ, and H_δ, which are located in the *visible* region (there are *ultraviolet* lines to the right of H_δ). The lines are seen to converge toward a point labeled H_∞ ("series limit"), beyond which the emission spectrum becomes continuous. This sequence of lines is now known as the **Balmer series.** It bears the name of the man who discovered a simple mathematical formula from which the wave lengths of all lines of this series can be computed. This achievement was of fundamental significance, since it showed that there is a simple functional connection between the different frequencies emitted by an atom and suggested that the emission of these frequencies is governed by a common physical law. In modern texts Balmer's relation is not written in terms of λ but rather in terms of the **wave number** $\tilde{\nu}$. $\tilde{\nu}$ is the reciprocal of the wave length; it indicates the number of waves per centimeter:

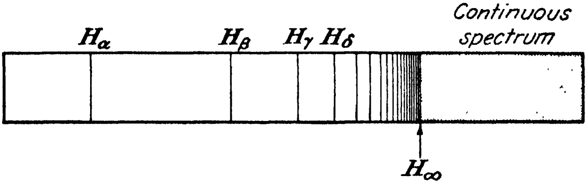


FIG. 35.1. The Balmer spectral series of hydrogen.

$$\tilde{\nu} = \frac{1}{\lambda} = \frac{\nu}{c}$$

where ν is the frequency and c the speed of light. In terms of $\tilde{\nu}$ **Balmer's formula** is as follows:

$$\tilde{\nu} = R_H\left(\frac{1}{2^2} - \frac{1}{n^2}\right) \tag{35.1}$$

R_H = **109,677.76 cm^{-1}** is the so-called **Rydberg constant** (for hydrogen) and n is an integer > 2. Placing $n = 3$ yields the wave number of the

H_α line, $n = 4$ yields $\tilde{\nu}$ for H_β, etc. For $n \to \infty$ we get the value of $\tilde{\nu}$ for the series limit

$$\tilde{\nu}_\infty = R_H\left(\frac{1}{2^2} - \frac{1}{\infty}\right) = \frac{R_H}{4}$$

The wave numbers computed from Balmer's formula are in excellent agreement with observed values.

Inspection of equation (35.1) suggests an interesting idea: Let us replace the term $1/2^2$ in the parenthesis by the term $1/1^2$. If we now compute values of $\tilde{\nu}$ for different integral values of $n > 1$, we obtain wave numbers corresponding to lines in the extreme ultraviolet region. Do the computed data correspond to an actually existing spectral series? Lyman succeeded in showing that they do (**Lyman series**).

Similarly, if we replace in equation (35.1) the term $1/2^2$ by $1/3^2$ and substitute consecutive integers for $n > 3$, we obtain wave numbers of lines in the infrared which were actually observed by Paschen (**Paschen series**).

It is thus apparent that equation (35.1) can be generalized to apply to all lines of the hydrogen spectrum:

$$\boxed{\tilde{\nu} = R_H\left(\frac{1}{k^2} - \frac{1}{n^2}\right)} \tag{35.2}$$

where both k and n are integers and $k < n$.

We obtain for $k = 1$, the Lyman series; for $k = 2$, the Balmer series; for $k = 3$, the Paschen series.

The above equation can also be written as follows:

$$\tilde{\nu} = \frac{R_H}{k^2} - \frac{R_H}{n^2} \tag{35.2a}$$

The wave number of any hydrogen line thus appears as the difference between two **spectral terms.** These terms are of the form R/k^2, where k is an integer.

This suggests that the wave numbers of some lines of a given series may be equal to the difference between wave numbers of two lines of another series. For example, consider $\tilde{\nu}_1$ and $\tilde{\nu}_2$, which represent the wave numbers of H_α and H_β of the Balmer series. Their difference is equal to the wave number $\tilde{\nu}_p$ of the first line of the Paschen series. This can be seen as follows. We get for the Balmer lines

$$\tilde{\nu}_1 = R_H\left(\frac{1}{2^2} - \frac{1}{3^2}\right); \qquad \tilde{\nu}_2 = R_H\left(\frac{1}{2^2} - \frac{1}{4^2}\right)$$

From this follows

$$\tilde{\nu}_2 - \tilde{\nu}_1 = R_H\left[\left(\frac{1}{2^2} - \frac{1}{4^2}\right) - \left(\frac{1}{2^2} - \frac{1}{3^2}\right)\right] = R_H\left(\frac{1}{3^2} - \frac{1}{4^2}\right)$$

The resulting expression is recognized as the wave number of a line of the Paschen series.

We have now enough observational material to be able to test a theory of the process of emission of light by the simplest atom, namely, hydrogen. We can now turn to the next step of our program and search for a suitable idea of an atomic model which could explain the emission of light according to the laws of mechanics and electrodynamics. Our theory should at least enable us to derive the equation (35.2) expressing the wave numbers of radiations emitted by hydrogen.

Thomson's Atomic Model

The road to a successful atomic model was a devious one. We can discuss only a few of the many proposed models which can be considered landmarks on the road to our contemporary ideas on atomic structure. The most important of the early attempts to conceive of a mechanical model of an atom was due to J. J. Thomson. He assumed an atom to be composed of positive and negative electricity, the former being associated with most of the atom's mass (see evidence discussed in Chap. 22). To the large, positively charged mass he assigned most of the volume of the atom. The simplest atom, *hydrogen*, was to consist, according to him, of a "drop" of positive electricity, about 10^{-8} cm in diameter (this dimension was based on deductions from the kinetic theory of gases) containing an equal amount of negative electricity in form of the much smaller electron in its interior. The equilibrium position of the electron was at the center of the positive spherical drop. Harmonic oscillations of the electron about this equilibrium point were supposed to cause emission of light.

Heavier atoms, capable of producing more highly charged ions than hydrogen, were supposed to contain a larger number of electrons inside a larger positive drop whose positive charge was to be equal to the combined negative charge of the electrons.

Thomson's X-ray Scattering Experiment

Thomson succeeded in estimating the number of electrons per atom for various elements from data on scattering of X rays. Let us suppose that we place a block of volume V containing N atoms of an element into the path of the X rays. According to the electromagnetic theory, the electromagnetic X-ray wave will force the electrons to vibrate inside the atoms. The vibrating electrons will emit X radiation in all directions. These will be the scattered X rays, which we observe.

Theoretical considerations enabled Thomson to obtain a formula relating the relative amount of X-ray energy scattered out of the original beam to the *number n of scattering electrons per unit volume.* Thus scattering observations led him to the determination of n. Knowing the number n of electrons per unit volume and the *number N of atoms per unit volume* (from Avogadro's number, the molecular weight, and the density), he was able to determine $Z = n/N$, the *number of electrons per atom. He found that, for the light elements, Z is about equal to half the atomic weight.* This would make the positive charge in a neutral atom (in electronic units) also about equal to half the atomic weight of the element.

Lenard's Electron Scattering Experiments

Experiments on scattering of various radiations by matter proved a very fertile method of "feeling out" the constitution of atoms. Lenard observed that the ability of electrons to penetrate matter increases with the speed of the electrons. He interpreted the elimination (by deflection) of *very fast* electrons from the original beam to be the result of a direct collision with an obstacle, a "force center." He imagined these force centers to cast an electron shadow, so to speak. From the proportion of deviated fast electrons he was able to estimate the "cross section" of the "impenetrable matter" contained in a given volume of a body. **He found the surprising result that the space occupied by matter is mostly "empty,"** *i.e.*, penetrable. Thus, for instance, if the "impenetrable centers" of a cubic meter of platinum could be compressed into a lump, they would fill the volume of a pinhead.

Rutherford's Alpha Scattering Experiment

Rutherford was using heavier projectiles than Lenard in his scattering experiments, which were designed to check the validity of Thomson's atomic model. Relatively simple considerations, based on electrostatics and on statistics, enabled Rutherford to deduce an expression for the relative number of α particles† (very fast-moving positive ions of doubly ionized Helium) deflected from their original direction of motion through a certain angle. The degree of agreement between the predicted amount of scattering and the actually measured value was to decide the fate of Thomson's atomic model.

Figure 35.2 shows the deviation of an α particle due to collision with a Thomson atom. In the absence of the atom the α particle would have moved along the straight line AB. As it enters the atom, however, its path is curved outward, and at a great distance from the atom it moves again along a straight line (CD), which forms the angle of deviation θ

---

† These particles are emitted by some of the so-called *radioactive elements.* Sources of α particles will be discussed in Chap. 38.

with the original direction of motion. In this discussion we shall consider merely the interaction between the heavy positive drop of the Thomson atom with the much lighter α particle which passes through it. Passing through such a drop the α particle will be repelled from the center but not in inverse proportion to the square of the distance! In fact, the repulsive force will *diminish* as the α particle approaches the center and will be *zero at the center* of the drop. It is easy to see why it should be so. Figure 35.3 shows the α particle at a certain point on its way through the Thomson atom. The dotted circle indicates an imaginary division of the spherical positive drop into two parts: (1) a charged sphere inside the spherical surface of radius R_1 on which the α particle is located and (2) a hollow charged sphere of external radius R_2 on whose inside surface the α particle is located. The hollow sphere exerts no forces upon the α particle. (It was stated on p. 209 that the force exerted upon a charge in the interior of a hollow charged conductor is zero.) The repulsive force is entirely due to the charge inside the inner sphere of radius R_1 and is such as if the charge were concentrated at the center of the sphere and acting upon the α particle according to Coulomb's law. The closer the α particle comes to the center the smaller is R_1 and, hence, also the repulsive force of the "inner charge." Hence, *an α particle passing very close to the center should go through a Thomson atom almost without deviation.*

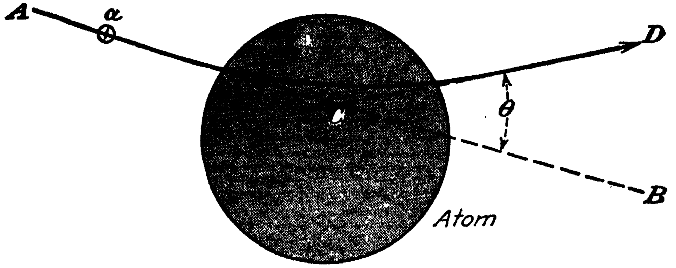


Fig. 35.2. Deviation of an α particle by collision with a Thomson atom.

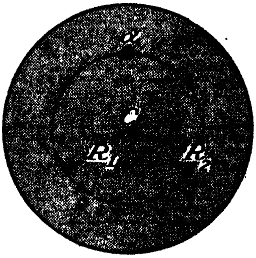


Fig. 35.3.

On the basis of such electrostatic considerations, taking into account the possibility that α particles may be deviated as a result of many collisions on their way through a thick foil, Rutherford derived a statistical formula giving the percentage of incident particles which would be deviated through a given angular range. A comparison between the figures predicted by this theory and observational data was to test the correctness of Thomson's ideas on atomic structure.

The apparatus of Rutherford, Geiger, and Marsden designed to check the above theory by observation of α scattering is shown in Fig. 35.4. R is a source of α particles which impinge upon a foil F. (The solid rays indicate the incident alphas and the dotted ones the scattered particles.) The α particles are detected with the aid of a zinc sulfide screen S. An incident particle causes the screen to emit a flash of fluorescent light, which enters the objective of the microscope M. The position of the microscope can be changed so as to detect particles scattered through different angles. The space through which the α particles have to pass is evacuated.

This simple method of registering individual atomic events, namely, the arrival of individual alpha particles, yielded results of crucial importance. The theory based on the Thomson atom predicted, for instance, that the most probable angle of scattering for alpha particles passing through a gold foil of 5×10^{-5} cm thickness would be the angle of 1°, scattering through larger angles being much less probable. For example, only one in 10^{13} particles was expected to be deviated through 30°, which is a vanishingly small fraction. *The above-mentioned experiments showed, however, a much larger scattering than was expected.* Angles of deviation of as much as 90° were observed! One in 8,000 alpha particles was scattered by a thin platinum foil through angles larger than 90°! Some of the scattered particles were even "reflected" back toward the source.

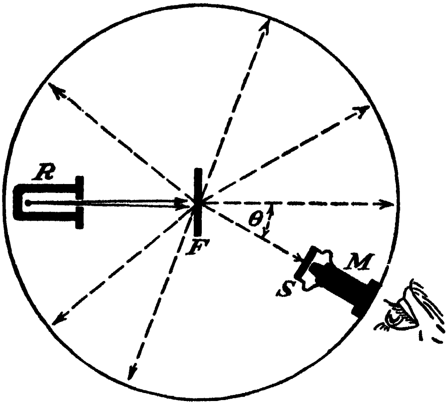


FIG. 35.4. Rutherford's scattering experiment.

Rutherford's Atomic Model

These results differed so radically from the predictions based on Thomson's atomic model that this model had to be rejected as being entirely wrong. The experimental results indicated in what way Thomson's atomic model had to be modified in order to explain the observations. The main weakness of Thomson's model consists in the fact that the repulsive force which is exerted upon the alpha particle is not large enough to cause as large deviations as are actually observed. *Rutherford realized that the desired strong repulsive forces would come into play if we imagine the entire positive charge, which in Thomson's model was spread over the entire atomic volume, to be concentrated in one point at the center of the atom, in the so-called nucleus.* Figure 35.5 shows for comparison Rutherford's positive nucleus N inside Thomson's positive drop (dotted circle). The interaction of an alpha particle with Rutherford's nucleus is quite different from its interaction with the Thomson atom. The force of interaction is given in Rutherford's model by Coulomb's law,

$$F = Qq/R^2$$

(where Q is the charge of the nucleus and q of the alpha particle). We see that *this force tends to become infinite as the distance R between the alpha particle and nucleus approaches zero.* Thus Rutherford's model enables us to account for very large deviating forces by assuming a sufficiently close approach between the alpha particles and the nucleus. A simple computation reveals that the trajectory of the alpha particle under such conditions is a hyperbola (solid curve in Fig. 35.5).

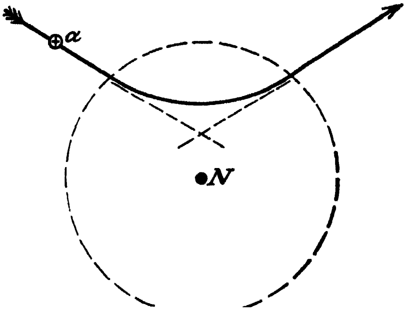


FIG. 35.5. Deviation of an α particle by collision with a Rutherford atom.

Replacing the Thomson model by his own nuclear atomic model, Rutherford derived a scattering formula which predicted the relative amount of scattering through various angles in good agreement with experiment. We shall accept this agreement between prediction and experiment as a confirmation of the correctness of Rutherford's picture of the atom and shall attempt now to estimate the size of the nucleus from results of scattering experiments.

Estimate of the Nuclear Radius. Let us consider a central impact between an alpha particle of charge $2e$ and a heavy nucleus of charge Ze. In this process the alpha particle will be reflected by the nucleus to where it came from. (This is a special case of scattering, namely, the case of a deviation through $\theta = 180°$.) We shall assume the nucleus to be much heavier than the α particle, so that it can be assumed to remain approximately at rest. If the kinetic energy of the alpha particle is $KE = \frac{1}{2}mv^2$ and if the mass of the atom is $M \gg m$, how close will the alpha particle come to the nucleus? It will move toward the nucleus until all its kinetic energy has been converted into potential energy. Then at the point of closest approach (let us say when the distance between the centers of the particles becomes R_0) we can equate the potential energy† which the α particle has acquired to the kinetic energy which it has lost by the time it reaches this point:

$$PE = \frac{(Ze)(2e)}{R_0} = \frac{1}{2}mv^2 = KE \qquad (a)$$

Solving for R_0 we obtain

$$R_0 = \frac{2Ze^2}{\frac{1}{2}mv^2} \qquad (b)$$

Using alpha particles of a radium C' source for which the kinetic energy of the alpha particles is known, Chadwick found that the number of particles scattered backward from a gold foil agreed well with the predictions of the Rutherford formula. This shows that Coulomb's law, on which Rutherford's formula is based, holds down to the distance of closest approach which is reached in head-on collisions between the alpha particles and gold atoms. Since Coulomb's law holds down to that distance, equation (b) must be valid. Substituting the data for $\frac{1}{2}mv^2$ of the Ra C' alpha particles and the nuclear charge Ze of the gold atoms in equation (b), Chadwick obtained $R_0 \simeq \mathbf{3 \times 10^{-12}}$ **cm** as the distance of closest approach. This distance can be interpreted as the *upper limit of the nuclear radius*, which we shall define as the distance at which Coulomb's law ceases to hold.

Nuclear Charge. But how did Chadwick know the value of Z for gold? The nuclear charge Ze enters Rutherford's scattering formula. Thus by measuring the percentage of the incident alpha radiation scattered in a given direction from foils of different materials, the *nuclear charge* Ze of various elements was determined. Geiger and Marsden found that *for light elements* $Z \simeq A/2$, *where* A *is the atomic weight.*

This result is in good agreement with Thomson's estimate of the number of electrons per atom from his X-ray scattering experiments. This result is to be expected, since in a neutral atom the positive charge of the nucleus must be equal to the negative charge of the electrons.

From a comparison between the values of Z obtained for various elements from α scattering experiments and the sequence number (atomic number) of the elements in Moseley's "natural system of elements," it was found that the two numbers are identical. We can thus state the following important result of an exhaustive survey of nuclear charges:

---

† The potential energy of a system of two point charges Q and q R cm apart is $PE = Qq/R$.

The nuclear charge of an atom is equal to its atomic number. (The nuclear charge is measured in units of e, where e is the charge of the electron.)

The Planetary Atomic Model. Let us now consider the distribution of the electrons in the atom. We may assume that they surround the nucleus. But how are they prevented from crashing into the nucleus under the influence of the Coulomb attraction? To answer this question Rutherford adopted a suggestion of Nagaoka. He assumed that electrons are kept from falling into the nucleus by the same mechanism which prevents the moon from falling onto the earth. Rutherford pictured the electrons as revolving about the nucleus, much as the planets do about the sun.

This picture of revolving electrons looks quite familiar to us. In our attempts to explain magnetism we assumed the existence of circular "molecular currents." Barnett's experiment appeared as a brilliant confirmation of the existence of whirling electrons. It is a great comfort to find that we arrive at the same picture independently through considerations of an entirely different kind. But we must not rejoice prematurely! We shall see later that such coincidences do not necessarily confirm a theory irrevocably.

Let us look now at some of the difficulties of Rutherford's *planetary atomic model*. First of all, a *stable planetary atom is inconsistent with Maxwell's electromagnetic theory*. A revolving electron is constantly subject to a centripetal acceleration and should, according to Maxwell, emit electromagnetic waves like any other accelerated charge. Thus a planetary electron should constantly lose energy by radiation, coming closer and closer to the nucleus in this process, finally crashing into it. Second, such an atom would not emit a line spectrum. An electron revolving at constant speed would emit a radiation of a frequency equal to the frequency of revolution. Now computation shows that, as the electron moves closer to the nucleus (on a spiral orbit as a result of radiative energy losses), its frequency of revolution increases. Thus radiation of steadily increasing frequency should be emitted by an electron spiraling toward the nucleus, giving rise to a continuous spectrum rather than to a line spectrum. Such difficulties suggest that *the existence of stable electronic orbits and the emission of a discrete line spectrum could not be accounted for on the basis of classical Newtonian mechanics and Maxwell's electrodynamics*. In order to achieve an explanation of the phenomena of emission and absorption of light and X rays on the basis of Rutherford's planetary atomic model, radically new postulates had to be adopted as a foundation of a new atomic mechanics. The way toward the necessary new assumptions was pointed out by results of previous research in two branches of radiation physics: the studies of heat radiation and of the photoelectric effect.

Planck's Quantum Hypothesis

The energy radiated at a given temperature by a hot "black body" (a body which absorbs all incident electromagnetic radiations) in different regions of the spectrum had been carefully studied. The spectral energy distribution curves obtained empirically are exemplified by the curve shown in Fig. 35.6. In this diagram the energy E emitted by the black body per unit wave length interval (*e.g.*, between 4,000 A and 4,001 A) per second is plotted in arbitrary units as a function of the wave length λ.

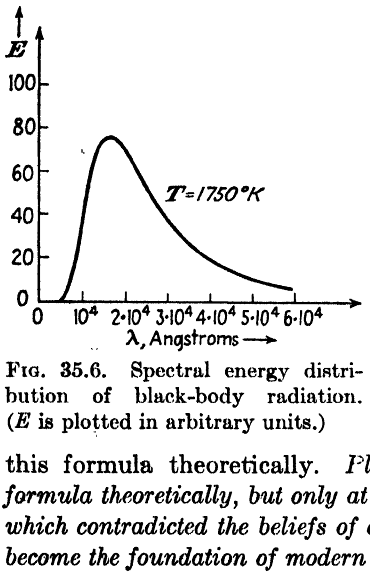


FIG. 35.6. Spectral energy distribution of black-body radiation. (E is plotted in arbitrary units.)

Attempts were made to find a general law which would represent the energy emitted as a function of λ at different temperatures.

Planck found by trial a functional relationship of the form $E = f(\lambda, T)$, representing the energy emitted as a function of wave length and temperature, which agreed very well with experimental findings for all values of λ. He then tried to determine what assumptions about the process of radiation had to be made in order to deduce this formula theoretically. *Planck succeeded in deriving his radiation formula theoretically, but only at the price of adopting a radical assumption which contradicted the beliefs of classical physics and which was destined to become the foundation of modern atomic physics.*

Planck considered the various frequencies in the continuous spectrum of the radiating "black body" to be emitted by oscillating electrons forming an assembly of harmonic oscillators of all conceivable frequencies. According to Maxwell a harmonically oscillating electron should continually lose energy by radiation and gradually diminish its amplitude. It would vibrate similarly to a pendulum swinging in sirup. Conversely, it should be possible to increase the amplitude (and thus the energy) of the oscillating electron by making it gradually absorb radiation at a certain rate. Such continuity of the processes of absorption and emission is demanded by the laws of classical physics. A pendulum of a given natural frequency could swing with any desired amplitude.

Planck had to give up the idea that a harmonic oscillator of a given natural frequency ν *could vibrate with any desired amplitude and that it could change its amplitude gradually from one value to another by gradual absorption or loss of energy! In order to deduce his law of heat radiation he had to assume that* **the oscillator could possess only certain stable values of energy and that it could change its energy only in discontinuous steps, gaining or losing the amount** $h\nu$ **of energy** (*quantum of energy*). *The stable energy states of the oscillator which correspond to oscillator energies* 0, $h\nu$, $2h\nu$, *etc., are characterized by the fact that* **the electron does not lose energy by radiation while vibrating with these energies.** h *is a fundamental constant of nature, the so-called* **Planck's constant** *or* **Planck's quantum of action:** $h = 6.62 \times 10^{-27}$ **erg-sec.** *It has the dimension of energy times time ("action," see Chap.* 10).

It appears thus that an oscillator of a given natural frequency can emit or absorb energy only in quanta $h\nu$, *which can be considered as "atoms of energy" peculiar to that particular oscillator. There is, however, no universal atom of energy, since the quantum* $h\nu$ *is proportional to the oscillator frequency and since oscillators of all conceivable frequencies can occur in nature.*

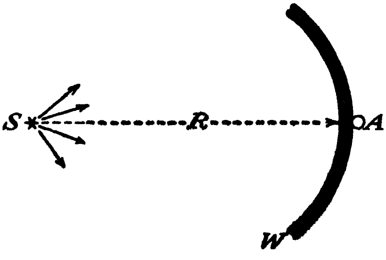


FIG. 35.7.

The process of emission of radiation was assumed to be discontinuous. The oscillator is supposed to change its energy suddenly from $n(h\nu)$ to $m(h\nu)$, in which process the energy difference $(n - m)h\nu$ is emitted in form of an electromagnetic wave. We can imagine for the time being this emitted wave to spread as a spherical electromagnetic pulse wave from the radiating atom.

How can we imagine now the process of absorption? Here a very serious difficulty is encountered if we try to picture a discontinuous process of absorption while adhering to the above picture of the emission of radiation. Consider, for instance, an atom on a distant star (S in Fig. 35.7) emitting an energy quantum, which is absorbed by an atom A in a terrestrial laboratory. If the energy of the emitted quantum is uniformly distributed over a spherical shell which has spread uniformly in all directions from S, it will be localized on the surface of a huge sphere of radius R, which light takes thousands of years to traverse. (Only a small portion W of the spherical wave front is shown in Fig. 35.7.) It would be quite inconceivable that the energy which has spread over the enormous span of the spherical wave front in the course of thousands of years could be gobbled up by the absorbing atom in a tiny fraction of a second!

Einstein's Photon Theory of Radiation

To find a way out of this difficulty Einstein took a bold step which at first glance looks like a step backward, since it appears to return us to Newton's corpuscular theory of light which we abandoned. *Einstein assumed that light (visible as well as infrared and ultraviolet), X rays, and in fact any electromagnetic radiation emitted by atomic constituents are not emitted in the form of spherical waves but rather in the form of packages of energy which he called* **photons.** We have no clear idea of what the photons "look like." We imagine them to be emitted in a definite direction with a definite momentum in a similar way to a particle. The absorption process is, similarly, pictured like the absorption of a particle, so that the difficulty discussed in connection with Fig. 35.7 disappears.

Einstein assumed the **energy of a photon** to be given by the expression

$$\boxed{E = h\nu} \qquad (35.3)$$

(where h is Planck's constant, ν the frequency, and E the energy in ergs) and the **momentum of the photon** by

$$\boxed{p = \frac{h\nu}{c}} \qquad (35.4)$$

(where c is the speed of light in a vacuum).

Equation (35.4) follows from equation (35.3) as follows: According to Einstein's law [equation (10.7)] the energy $h\nu$ is equivalent to the mass $m = h\nu/c^2$, which we shall designate as the mass of the photon. We can write $\underline{h\nu} = (mc)c = \underline{pc}$ where the product $mc = p$ is the momentum of the photon. This is the desired equation (35.4).

We infer from equation (35.3) that **the energy of a photon increases as the wave length decreases.**

But can we consider the difficulty of explaining the absorption of a light quantum as a good enough justification for abandoning the wave theory of light, which has correlated most of our knowledge of light so outstandingly well? The answer to this question is that Einstein had additional supporting evidence which will be considered below for his hypothesis and that he did not abandon the wave theory of light but rather supplemented it and reinterpreted the meaning of the concept of wave amplitude. Einstein attempted to correlate the corpuscular and wave-theory conceptions of light in the following fashion: *According* *to Maxwell's electromagnetic theory the intensity of a light wave at a given point is measured by the square of the amplitude of the wave, i.e., the square of the maximum electric or magnetic field strength which is observed at the point.* **According to Einstein the square of the amplitude of the electric or magnetic field strength measures the photon density,** *i.e., the number of photons per unit volume at a given location.* Thus, whereas the electromagnetic theory pictures the diminution of illumination with distance from a source as being due to a diminution of the amplitude of the electromagnetic field, the photon theory describes it as being due to "dilution" of the emitted photons over a wider space.

The Photoelectric Effect

Let us now examine the main experimental evidence which led Einstein to the formulation of the photon hypothesis. Lenard was investigating a phenomenon which was discovered by Heinrich Hertz as a by-product in the course of his experiments with electromagnetic waves. This phenomenon (**photoelectric effect**) consists of emission of electrons from metal surfaces which are illuminated by electromagnetic waves of high frequency, such as X rays and ultraviolet, visible, and infrared light. This effect is most pronounced with X rays and short-wave ultraviolet light, which yield photoelectrons with practically any metal surface. Visible light shows the photoelectric effect with alkali and alkaline-earth metals. A few substances, such as cesium oxide, emit photoelectrons when illuminated with the near infrared radiation. *All substances possess a definite* **long-wave-length limit** λ_0 *such that only light of* $\lambda < \lambda_0$ *will cause emission of photoelectrons.*

Figure 35.8 shows schematically an arrangement for the study of the photoelectric effect. Light *L* impinges upon a metal plate *C* inside a highly evacuated tube *V*. The photoelectrons released from *C* allow a current to flow between the electrodes *A* and *C*, thus closing the circuit of the galvanometer *G*. The "voltage divider" *P* allows a variable *retarding* voltage to be applied between the plates *A* and *C*, such that only electrons having more than a certain minimum of kinetic energy will be able to reach plate *A*. This arrangement enables one to determine the maximum kinetic energy which occurs among the electrons emitted from *C* by simply increasing the retarding voltage until the galvanometer current drops to zero.

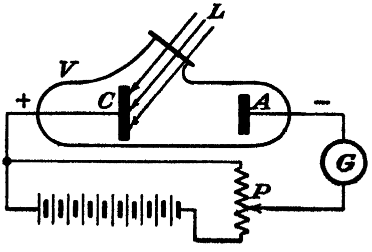


Fig. 35.8.

Lenard found a baffling result with an experimental setup of this type. According to the electromagnetic theory, one would expect the speed of the electrons emitted by C to increase with the intensity of the illuminating light. One might imagine that electrons are thrown out of the metal by the impulse which is imparted to them by the electric field of the light wave. Since an increased intensity of light is associated with an increased electric field amplitude, according to Maxwell's theory of light, the momentum acquired by electrons under more intense illumination should be greater. *Lenard found, however, that the intensity of light had no effect whatsoever on the velocity distribution of the photoelectrons!* At any given illumination various speeds occurred among the emitted electrons ranging from zero to a certain *sharp maximum velocity.*

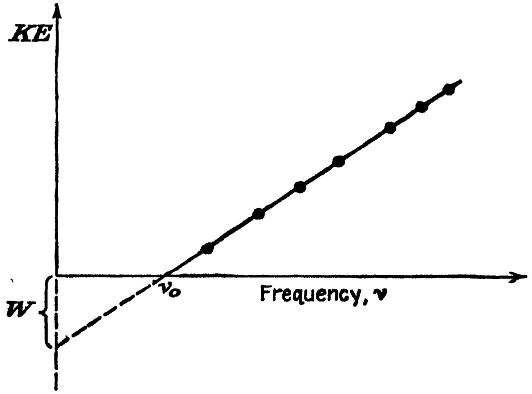


FIG. 35.9. Maximum kinetic energy (KE) of photoelectrons as a function of the frequency (ν) of the radiation. (KE and ν are plotted in arbitrary units.)

The maximum velocity remained the same regardless of how many thousand times the light intensity was augmented. *Increased illumination merely increased the total number of emitted electrons.* This was incomprehensible on the basis of the classical electromagnetic theory.

The situation was made even more puzzling by Lenard's discovery that the maximum kinetic energy of emitted electrons could be influenced by the color of the illuminating light. He found in experiments with monochromatic radiations that *the maximum speed of photoelectrons increased with the frequency of the illuminating radiation.*

Figure 35.9 shows Lenard's results. KE is the *maximum* kinetic energy of the photoelectrons, and ν is the frequency of the light used. We see that no electrons are emitted (KE = 0) until the frequency of light reaches the minimum value of ν_0 (**photoelectric threshold**). From then on KE increases in direct proportion to the frequency. Thus the relation between KE $= \frac{1}{2}mv_0^2$ (v_0 = maximum electron speed) and ν is represented by the equation of the straight line of Fig. 35.9:

$$\tfrac{1}{2}mv_0^2 = a\nu - W \tag{35.5}$$

(a is the slope of the line and $-W$ its intercept with the vertical axis.)

Einstein's Photoelectric Equation

Einstein recognized the physical significance of Lenard's equation (35.5). All terms of this equation must have the same physical dimension, namely, that of energy. The term $a\nu$ reminds us immediately of $h\nu$, the energy of a photon. Einstein showed that the value of the slope of the line in Fig. 35.9 is equal to Planck's constant. Setting $a = h$ in equation (35.5) we obtain Einstein's famous photoelectric equation:

$$\boxed{\tfrac{1}{2}mv_0^2 = h\nu - W} \tag{35.6}$$

Einstein interpreted this equation as follows:

A photon of energy $h\nu$ strikes the metal surface. Its entire energy is imparted to an electron. If the electron were not subjected to retarding forces, it would escape with a speed given by the equation $\tfrac{1}{2}mv^2 = h\nu$. But the electron has to overcome a retarding field as it leaves the metal surface (think, for instance, of attraction between a charge and a neutral metal body) and thus loses W ergs of kinetic energy. It thus escapes with a kinetic energy KE $= h\nu - W$. W is called the **work function.**

As we diminish ν, we eventually reach the value of $\nu = \nu_0$ for which KE $= 0$. This makes $W = h\nu_0$. This means that the work function can be determined by finding the threshold frequency ν_0 below which no photoelectrons are emitted. *The value of Planck's constant can be determined from the slope of the line in Fig.* 35.9.

The above theory explains why it is the *frequency,* rather than the *intensity,* of light that determines the maximum energy of photoelectrons and why this energy is directly proportional to the frequency. We have only to explain now why all velocities from v_0 down to 0 occur among the emitted electrons. Light penetrates through many superficial layers of atoms, so that many of the electrons ejected from the depth are likely to suffer repeated collisions with atoms before they escape, thus losing various amounts of energy, in some instances losing enough to be prevented from leaving the metal surface.

Einstein also proposed a second equation:

$$n = \frac{E}{h\nu} \tag{35.7}$$

E is the energy of the absorbed electromagnetic radiation, and ν is its frequency. Since $h\nu$ is the energy of a photon of this monochromatic radiation, n gives the number of absorbed photons. n is the maximum number of photoelectrons which could be released theoretically from a metal surface of zero work function. Actually, the number of emitted photoelectrons is very much smaller than this theoretical maximum value for reasons which have been mentioned above.

The importance of equation (35.7) lies in its application, not to the photoelectric effect, but rather to photochemistry. Einstein's fundamental law of photochemistry states that **each absorbed light quantum causes one molecule to undergo a primary reaction.** Equation (35.7) enables one to determine the number of the "activated" molecules.

The Law of Duane and Hunt

If the energy $h\nu$ of the photon is very great as compared with the work function W, we can neglect W in equation (35.6) and write

$$\tfrac{1}{2}mv_0^2 \cong h\nu \tag{35.8}$$

This is a good enough approximation for X-ray photons. This equation states that an X-ray photon of energy $h\nu$ will emit electrons with velocities no greater than v_0.

This equation admits also of the following interpretation: Fast electrons of velocity v_0 produce X rays when they are stopped by a target. When an electron is stopped by a single collision, it loses the energy $\frac{1}{2}mv_0^2$, which we can imagine to be converted into the energy of a photon of $h\nu^*$ ergs. If the electron's velocity is not reduced to zero by the collision, it loses less than $\frac{1}{2}mv_0^2$ of energy, the energy of the emitted photon is less than $h\nu^*$, and hence, its frequency is less than ν^*.

The equation

$$eV = h\nu^* \tag{35.8a}$$

has been introduced before [equation (34.3)] as the **law of Duane and Hunt.** For this equation h is Planck's constant, e is the electronic charge, and V is the potential difference through which an electron has to fall in order to gain the kinetic energy $\frac{1}{2}mv_0^2$. Thus *the above equation enables one to compute the maximum frequency (and hence the shortest wave length) of X rays emitted by an X-ray tube which is operated at the voltage V.*

Evidence for the Existence of Photons

The success of the photon hypothesis in explaining a great variety of phenomena stimulated experimental research designed to confirm the existence of photons. We shall cite below some of the accumulated direct experimental evidence confirming the photon hypothesis.

1. **Absence of Anticipated Delay in Electron Emission.** Suppose we illuminate the surface of a metal whose work function is, say, $W = 6.63 \times 10^{-12}$ erg. This means that we have to impart to an individual electron at least 6.63×10^{-12} erg of energy in order to enable it

to leave the metal surface. According to the electromagnetic wave theory of light, energy can be considered as pouring continually into an atom, which absorbs the radiation. The amount of energy E which enters the atom per second is equal to the cross-sectional area A of the atom times the energy flux per unit area ϕ (intensity of illumination). $E = A\phi = (\pi R^2)\phi$ (where R is the radius of the atom).

Let us assume that $\phi = 20$ ergs/(cm^2)(sec) and that the atomic radius $R = 1.5 \times 10^{-8}$ cm. Then the energy absorbed by the atom per second is $E = \pi(1.5 \times 10^{-8})^2 \times 20 = 14.13 \times 10^{-15}$ erg/sec. *How long will it take the atom to accumulate* $W = 6.63 \times 10^{-12}$ *erg of energy, which is sufficient to allow one electron to escape?* The answer is simply

$$\underline{t} = \frac{6.63 \times 10^{-12}}{14.13 \times 10^{-15}} = \underline{469 \text{ sec}}$$

This means that we should have to wait at least 469 sec after switching on our source of light before the first electron is emitted.

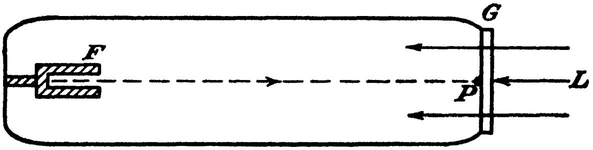


Fig. 35.10.

Experiment shows, however, that this is not at all what happens! The first electron may be emitted after 1, 10, or 10^{-3} sec or less! The only way to understand this result is by assuming that energy reaches the metal surface in packages, so that radiation is not uniformly distributed over the surface but, rather, is very highly concentrated at those points which happen to be struck by a photon. (Naturally the frequency of the light used must be larger than the threshold value ν_0.)

2. Recoil of Radiating Atoms. According to the classical electromagnetic theory, a radiating atom emits a spherical wave so that, for reasons of symmetry, the emitting atom should not suffer a recoil. According to the photon theory, on the other hand, the atom should suffer an eastward recoil when it emits a photon of momentum $h\nu/c$ in the westward direction. The momentum of the atom should be equal to the momentum of the photon according to the law of conservation of momentum.

The following experiment should enable us to decide which one of the two predictions (and hence, which one of the two theories) is the correct one (Fig. 35.10): Atoms escape in a narrow beam from the hole in a furnace F, which is placed in a highly evacuated tube. They impinge upon the glass window G at point P and are deposited there, forming a small black spot. If light L is now allowed to enter the vacuum tube parallel to the atomic beam, we shall expect the atoms to absorb and to reemit radiation (*fluorescence*) along the path between F and P. According to the wave theory of emission, the process of emission and absorption of radiation in this symmetrical case should not deviate the atoms from their path and hence the spot of deposition P should remain as sharp with the light source on as it was in darkness. According to the photon theory, however, photons will be absorbed by the moving atoms and reemitted in all directions; in particular, a photon could fly off at right angles to the atomic beam, in which case the emitting atom would suffer a recoil in the opposite direction and land far away from the spot P. *The photon theory leads us to expect scattering of the atomic beam as a result of the fluorescence of the atoms in the beam.* The greatest deviation to be expected is for atoms which emit a photon at right angles to the beam at the moment when they leave the furnace F.

Experiments performed with sodium atoms illuminated with a monochromatic source of sodium D light (Na atoms absorb and reemit radiation of this frequency) confirmed quantitatively the predictions of the photon theory.

3. **Compton Effect.** Compton found that a certain type of X-ray scattering can be understood only if we consider the scattering process as an elastic collision between a photon and an electron. The details of Compton's experiment will be considered in Chap. 37. Compton's results may be considered the most convincing evidence in favor of the photon theory of light.

The Dual Nature of Light

At this point the student must be thoroughly confused. We spent much space and mental effort in establishing convincing evidence for the wave theory of light, and now we have demonstrated the particle nature of light. Which view is the correct one?! The answer is that both are correct. When we call a photon a particle, we merely establish an analogy between a photon and, say, a moving pebble which we can easily observe. *But there is no reason in the world why a photon should resemble an object familiar to us!* Similarly, when we speak of light waves, we are justified only in implying that we are dealing with a periodic process in space and time which in many respects resembles waves in a material medium, but we have no right to assume that this analogy is a complete one.

Our experience with light can be summarized by stating that it is a process which cannot be completely described in terms of the wave theory alone or *of the corpuscular theory alone. Both theories emphasize different aspects of the same phenomenon. When we are observing the passage of light through constraints such as slits, we have to resort to the wave theory to interpret the distribution of the light energy in space, and when we discuss phenomena involving emission and absorption of light (such as the photoelectric effect, fluorescence, etc), we rely on the photon theory for the interpretation of our observations.*

APPENDIX

Practical Applications of the Photoelectric Effect

A photoelectric cell consists, in its simplest form, of two electrodes contained in an evacuated glass tube. The negative electrode is usually in the shape of a semicylinder, and the positive electrode is a thin rod located along the cylinder axis. The cathode (semicylinder) is coated with a material of low work function (such as cesium oxide) if the cell is to be used with visible light.

The simplest application of the photocell is as a "light switch." When dark, it constitutes a break in a d-c circuit, but when light falls upon the semicylinder, the cathode emits electrons which flow to the positive rod so that the cell becomes a conductor; *i.e.*, the light switch is closed. If actuated by invisible light (ultraviolet or infrared), such a light switch can be used as a burglar alarm.

The fact that the photoelectric current is proportional to the intensity of illumination makes the cell usable as a *photometer*. A photocell is an essential part of **spectrophotometers,** which are used extensively in biochemical analysis. Such instruments determine the absorption spectrum of a solution by sending light of varying wave lengths through it and catching the transmitted beam on a photoelectric cell, where a measurable electron current is generated.

The photoelectric effect is also used in television in the so-called Iconoscope tube which contains a screen covered with tiny metal particles acting like miniature photoelectric cathodes. The principle of television is not difficult to grasp. The student is referred to textbooks on electronic engineering for reference.

QUESTIONS AND PROBLEMS

1. Under the influence of what forces do the electrons in Thomson's model maintain a stable configuration?

2. Is stability of the electrons in Thomson's model essential to their ability to emit electromagnetic radiation?

3. Compute (*a*) the second longest wave length of the Balmer series, (*b*) the third longest wave length of the Lyman series.

4. Compute the wave length of the "series limit" of the Paschen series.

5. A line of $\lambda = 4{,}102.9$ A is found in the spectrum of hydrogen. Identify the series to which it belongs.

6. How do we know that the deviation of alpha particles in Rutherford's scattering experiments is due to interaction of the alpha particles with the positive charge of the atoms and not with the electrons?

7. Suppose you determine the number of alpha particles scattered backward by Rutherford's method (Fig. 35.4) and find that at low speeds of alpha particles the measurements agree with the predictions of the Rutherford formula but that at high speeds they do not. What conclusion would you draw from these results about the nature of nuclear forces?

8. Compare two quanta of electromagnetic radiation which are specified by the wave lengths $\lambda_1 = 5{,}000$ A (visible light) and $\lambda_2 = 0.1$ A (X rays).

a. What are the energies of these quanta?

b. How large are their momenta?

c. How fast are the fastest electrons ejected by these quanta from a metal of work function $W = 5 \times 10^{-12}$ erg?

d. How fast are the slowest photoelectrons?

9. Suppose Planck's constant were $H = 6.62 \times 10^{8}$ erg-sec instead of

$$h = 6.62 \times 10^{-27} \text{ erg-sec}$$

Discuss how a pendulum struck by a hammer would behave under these conditions.

10. In what way does Planck's hypothesis constitute a radical departure from classical physics?

***11.** Can you deduce a law relating the number of emitted photoelectrons (*a*) to the intensity of the illuminating light? (*b*) to its frequency? If you can't do it for visible light, can you do it for X rays?

12. Could visible radiation be conceivably emitted by a metal target of work function $W = 9 \times 10^{-12}$ erg, bombarded with slow electrons of negligibly small speed? What would be the highest frequency of the emitted radiation?

13. Can you suggest an experiment, other than those mentioned in the text, for the purpose of testing the photon hypothesis?

14. A zinc plate is illuminated by ultraviolet light of 2,657 A. The intensity of illumination is 36.52 ergs/(cm^2)(sec). What will be the speed of the fastest ejected photoelectrons when the intensity of illumination is increased by a factor of 3?

15. Imagine an electron held in a stable equilibrium by forces of such a magnitude that it emits a radiation of the wave length $\lambda = 10{,}000$ A when allowed to oscillate. Compute the amounts of energy which this oscillator can possess according to Planck's original quantum hypothesis.

16. Compute the long-wave-length limit λ_0 and the threshold frequency ν_0 for a metal surface where work function is 5.0×10^{-12} erg.

17. The intensity of violet ($\lambda = 4{,}000$ A) illumination of a surface is given as 10^{-3} watt/cm^2. Express this illumination in terms of the number of photons impinging each second upon a unit surface.

CHAPTER 36

BOHR'S ATOMIC THEORY

a. Is Bohr's atomic model based on entirely new laws of physics, or are the laws of classical mechanics used in conjunction with additional postulates?

b. Which features of Bohr's atomic model are inconsistent with classical physics?

c. In what ways does Bohr's picture of a planetary atom differ from the Copernican picture of the solar system?

d. How would you define a "stationary orbit"?

e. How does Bohr picture the process of (1) emission, (2) absorption of radiation?

f. Is the emitted frequency of light equal to the frequency of revolution of the electron in its orbit? Is there a reason to expect that it should be so?

g. How does Bohr explain the emission of a spectral series?

h. How do you explain (1) fluorescence? (2) Stokes' law ("emitted light is never of a shorter wave length than the absorbed light")? (3) resonance fluorescence (reemission of light of the same frequency as that used for excitation)?

i. Why is the absorption spectrum of a gas different from its emission spectrum?

j. How can an atom be "excited"?

k. What does the Franck-Hertz experiment demonstrate?

l. Can an atom be excited by electrons of *any* energy? Can it be excited by photons of *any* energy?

m. How does Bohr's theory picture the emission of characteristic X rays?

n. How does Bohr's theory picture the mechanism of absorption of X rays?

Dynamics of Rutherford's Model of the Hydrogen Atom

Let us now examine Rutherford's atomic model in the light of physical theories (Planck's quantum theory and Einstein's photon theory) which have been discussed in Chap. 35.

According to Rutherford the electrons revolve about the nucleus, much as the planets do about the sun. We can derive simple expressions for the kinetic and potential energy of the planetary electrons, relating these magnitudes to the distance r of the electrons from the nucleus (assuming circular orbits for simplicity). We shall begin our considerations with the hydrogen atom, which has the atomic number 1, and hence, according to Chap. 35, must have the nuclear charge of one elementary unit and, therefore, only one planetary electron.

The potential energy of the revolving electron can be easily computed from the equation

$$\mathrm{PE} = -\frac{Qq}{r} \tag{36.1}$$

which follows from equation (15.7). It states how much work must be done on a charge $-q$ in moving it to infinity from a distance of r centimeters from the charge $+Q$. In the case of hydrogen, $|Q| = |q| = |e|$, and hence,

$$\text{PE} = -\frac{e^2}{r} \tag{36.1a}$$

For the sake of generality, we shall assume a nucleus of charge Ze at the center of revolution of a single electron. Then

$$\text{PE} = -\frac{(Ze)e}{r} = -\frac{Ze^2}{r} \tag{36.1b}$$

(The minus sign in the above equations is due to the fact that PE is proportional to the product of two *opposite* charges. It expresses the fact that there is attraction.)

The kinetic energy of the electron can be related to its potential energy through the following simple consideration: In order to perform a circular motion at constant speed in its orbit, the electron must be acted on by a centripetal force. The magnitude of this force is provided by the electrostatic attraction $F = (Ze)e/r^2 = Ze^2/r^2$ between the nucleus and the electron. The centripetal force, $F_c = mv^2/r$ is determined by the orbital velocity v of the electron and by the radius r of its orbit. Equating these two expressions for the centripetal force to each other we obtain

$$\text{Centripetal force} = \frac{Ze^2}{r^2} = \frac{mv^2}{r} \tag{36.2}$$

If we write this equation as follows,

$$\frac{1}{2}\frac{Ze^2}{r} = \frac{1}{2}mv^2 = \text{KE} \tag{36.2a}$$

i. e.,

$$\frac{1}{2}(-\text{PE}) = \text{KE}$$

we see that the absolute value of the kinetic energy is equal to one-half of the absolute value of the potential energy of the electron.

The total energy E of the electron is given by $E = \text{KE} + \text{PE}$. Substituting the values for KE and PE from equations (36.2a) and (36.1b) we get

$$\underline{E} = \frac{1}{2}\frac{Ze^2}{r} + \left(-\frac{Ze^2}{r}\right) = -\frac{Ze^2}{\underline{2r}} \tag{36.3}$$

The fact that the total energy is negative need not disturb us. This simply means that the potential energy of the planetary electron, which is negative (indicating an attractive force) by convention, is larger in absolute value than its kinetic energy. At infinity (for $r = \infty$) the potential as well as the total energy of the electron is zero. As the orbit of the electron shrinks to smaller values of r, the *absolute value* of the total energy increases whereas its *algebraic value* decreases.[†] *This means that energy must be supplied to an electron revolving in an orbit of radius r in order to transfer it to an orbit of larger radius.*

We thus obtain the following dynamic picture of Rutherford's hydrogen atom:

The electron revolves in an orbit of radius

$$r = \frac{Zc^2}{mv^2} \tag{36.2b}$$

[this follows from equation (36.2) by solving for r] with a frequency $f = v/2\pi r$. Its total energy is given by equation (36.3), and it possesses an angular momentum Ang. Mom. $= I\omega$, where $I = mr^2$ is the moment of inertia of the electron with respect to an axis passing through the nucleus and $\omega = v/r = 2\pi f$ its angular velocity. Substituting these values in the expression for the angular momentum we obtain

$$\text{Ang. Mom.} = I\omega = (mr^2)\left(\frac{v}{r}\right) = mvr \tag{36.4a}$$

Inadequacy of the Rutherford Model

Our objective is to devise an atomic model which would allow us to deduce from mechanical considerations the general formula for the frequencies of the hydrogen spectrum [equation (35.2)]. The above model does not seem to offer any hope of achieving this aim. It could explain the emission of one single frequency equal to the frequency of revolution of the electron if we are willing to assume that the energy emitted by the revolving electron in the form of electromagnetic waves is somehow replenished in the atom as fast as it leaks out. But then we should have to explain why all bodies don't constantly emit light at all temperatures. If, on the other hand, we do not assume a mechanism which supplies energy to the radiating electron, the latter will spiral into the nucleus with an ever-increasing frequency of revolution, emitting a spectrum of a continuous range of wave lengths. This is as much at variance with our observations as the prediction of one single frequency.

It is evident that, if we are to adhere to Rutherford's model, we must, in order to avoid the above difficulties, make a radical assumption which contradicts our experience with emitters of electromagnetic waves. **We must assume that the revolving atomic electrons do not emit electromagnetic radiation, although according to Maxwell's theory they should do so.**

[†] For example, -10 is algebraically smaller than -5.

That a revolving electron should emit electromagnetic waves in much the same way as an oscillating electron is apparent from the following consideration, which shows that a revolving body can be considered as performing simultaneously two vibrations which are perpendicular to each other and 90° out of phase. Consider Fig. 36.1. The body B revolves about the point O, and owing to parallel light coming from above, a shadow S_x of B is seen to move harmonically along the x axis, and similarly, owing to parallel light coming from the right, a harmonically moving shadow S_y appears on the vertical axis. We can read off from the diagram the expressions for the excursions of the horizontal and of the vertical shadows:

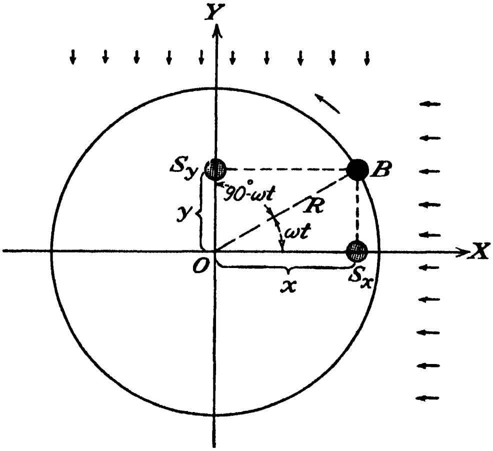


FIG. 36.1.

$$
\begin{aligned}
y &= R \sin \omega t \\
x &= R \sin (90 - \omega t)
\end{aligned}
$$

We see that the two harmonic motions are 90° *out of phase.* This phase difference between the motions is easily visualized with the aid of the diagram. At the moment S_x has reached its maximum displacement, S_y passes through zero, and conversely, as S_x approaches zero, y approaches its maximum value.

Let us imagine S_x to be an actual body oscillating along the X axis. How would we have to move it up and down (without disturbing its horizontal oscillation) so as to make it always coincide with the revolving sphere B? We should have to vibrate it up and down exactly as the shadow S_y is oscillating. Then the vertical as well as the horizontal excursion of this body (which is performing simultaneously two vibrations, a vertical and a horizontal one) will be the same as the coordinates of B. Thus a body vibrating along the X axis can be made to trace a circular path if it is at the same time performing a vertical oscillation of equal amplitude 90° out of phase with the horizontal one.

Bohr's Theory of the Hydrogen Atom

The realization that a revolving electron can be considered as a charge which performs two oscillations at the same time (which are mutually perpendicular and 90° out of phase) places it in the same class as a harmonic oscillator. It creates a connection with the ideas of Planck. We saw that, according to the quantum theory, an electronic harmonic oscillator does not emit electromagnetic waves except at the moment when it loses suddenly a quantum of energy $E = h\nu$ (where ν is the frequency of the oscillator). This assumption is as foreign to Maxwell's electromagnetic theory as the assumption of nonradiating revolving electrons. Since the revolving electron can be considered as a special type of harmonic oscillator, the idea suggests itself of applying Planck's quantum theory to Rutherford's atomic model. But if we are to explain the great variety of spectral frequencies emitted by the hydrogen atom, we cannot assume that the atomic oscillator retains a constant natural frequency ν while it emits successively quanta of energy $h\nu$. We rather assume that the emitted quanta of energy differ, depending on the initial and final energy of the electron. An electron of initial energy content E_a will be assumed to emit an energy quantum $h\nu_a$ of frequency ν_a, thus retaining a lower energy E_b. The next quantum of energy radiated will be, according to our assumption, emitted in the form of a photon of different frequency ν_b, so that the new "energy level" of the electron will be $E_c = E_b - h\nu_b$ (where $\nu_b \neq \nu_a$), etc.

In making this assumption we refrain from correlating the frequency of the emitted quantum with the frequency of revolution of the electron. *The emitted frequency is determined entirely by the difference between the initial and final energy levels of the electron.* Thus in the above example of transition from energy level E_b to level E_c the frequency ν_b is given by

$$\nu_b = \frac{E_b - E_c}{h}$$

Such a scheme would indeed be capable of yielding a variety of spectral frequencies if a suitable way were found to assign a sequence of energy levels to the electron. An appropriate scheme for the assignment of discrete energy values E_1, E_2, E_3, . . . to the planetary electron was suggested by Bohr. According to the classical theory the total energy of the planetary electron can assume an infinite number of values which are determined by equation (36.3) for any arbitrarily chosen value of r. Bohr made an assumption which restricted the possible values of the total energy to certain discrete amounts. This is equivalent to the statement that *the electron can revolve only at certain "allowed" distances from the nucleus.* His restrictive principle was based on Planck's quantum of action h, which has the same physical dimension as angular momentum. **Bohr assumed that the planetary electron could move without losing energy by radiation only in those orbits in which its angular momentum** mvr **[see equation** (36.4*a*)] **is an integral multiple** n **of** $h/2\pi$. This fundamental *postulate of Bohr* is expressed by the following equation:

$$\boxed{mvr = n\frac{h}{2\pi}} \tag{36.4}$$

This was an arbitrary assumption made by Bohr, justified only by the successful theoretical predictions (see page 729) to which it led. *This postulate, together with the postulated frequency condition*

$$\boxed{\Delta E = E_0 - E_f = h\nu} \tag{36.5}$$

form the core of Bohr's atomic theory.

The latter equation states that, whenever the planetary electron loses the energy ΔE, **thus reducing the atomic energy level from an initial value of** E_0 **to a final value of** E_f, **a photon of energy** $h\nu$ **is emitted.** *As to the converse process of absorption, only those photons can be absorbed by an atom for which the product* $h\nu$ *is equal to the energy difference* $E_0 - E_f$ *between the initial energy level* E_0 *of the planetary electron and a second, higher energy level* E_f *to which the electron is raised.*

Equation (36.5) can be written in the form

$$\boxed{\nu = \frac{E_0 - E_f}{h}} \tag{36.5a}$$

in which the emitted (or absorbed) frequency is expressed in terms of the energy levels.

The process of emission and absorption of light is thus explained in terms of a mechanical model by the jumping of electrons from one "allowed" orbit to another. If the final mechanical energy of the electron is lower than the initial energy, a light quantum is emitted; if it is higher, a photon is absorbed *in its entirety.*

The frequency condition (36.5) and the above-mentioned picture of the process of absorption of light can be considered to be a generalization of Einstein's theory of the photoelectric effect. Einstein assumed, in order to explain the ejection of electrons from matter, that the energy of a photon is imparted in its entirety to an electron. If this energy is high enough, the electron escapes and we observe the photoelectric effect. Bohr considered the interaction between light and isolated atoms in the case where the energy is insufficient to allow the electron to escape. He assumed that, instead of escaping from the atom, the electron merely escapes from its original orbit to be stored in a new orbit, provided the appropriate amount of energy is supplied to it. He generalized this picture further by considering the converse process, in which an electron jumps from a higher orbit into one of lower energy, the difference in energy between the two orbits being emitted in form of a photon.

Bohr's theory is based on a mixture of the laws of classical physics and of special assumptions which are alien to classical mechanics and electrodynamics, such as the quantum hypothesis of discrete energy levels and the assumption of certain stable orbits in which electrons can move without radiating energy. Bohr's postulates enable us now to deduce a law of radiation by combining them with the classical equations (36.2b) and (36.3). It remains to be seen how well the predictions of this law agree with experimental findings.

The radii of the allowed Bohr orbits can be obtained from equation (36.2) if we apply to it the restriction that the electron could have only angular momenta determined by Bohr's postulate expressed by equation (36.4):

$$\frac{e^2}{r^2} = \frac{mv^2}{Zr} \tag{36.2}$$

dividing equation (36.4) by mr^2 we obtain

$$\frac{v}{r} = n\frac{h}{2\pi mr^2} \tag{36.4b}$$

squaring equation (36.4b) we get

$$\frac{v^2}{r^2} = n^2\frac{h^2}{4\pi^2m^2r^4} \tag{36.4b'}$$

substituting (36.4b') in (36.2) we obtain

$$\frac{e^2}{r^2} = \frac{mr}{Z}\frac{v^2}{r^2} = \frac{mr}{Z}n^2\frac{h^2}{4\pi^2m^2r^4} \tag{36.6}$$

Solving for r, we get the following expression for the **radius of the allowed Bohr orbit:**

$$\boxed{r = \frac{n^2h^2}{4\pi^2me^2Z}} \tag{36.6a}$$

(h = Planck's constant, e = electronic charge, m = electronic mass, Z = atomic number.)

The smallest radius is obtained for $n = 1$. The atom is said to be in its **ground state** when the electron circles in this orbit. Since all symbols on the right side of equation (36.6a) designate constants except n, which stands for a variable *integer*, only a discrete set of orbital distances exists. The orbital radii are seen to increase as the square of the so-called **principal quantum number** n.

We can now compute the *values of the atomic energy levels.* Equation (36.3) is valid for any value of r. We now introduce the restriction that there is only a discrete set of possible orbital radii for the electronic orbits. By substituting the value for r from equation (36.6a), which determines the "allowed" orbital radii, into equation (36.3) we obtain

$$E = E_n = -\frac{Ze^2}{2}\frac{1}{r} = -\frac{Ze^2}{2}\frac{4\pi^2me^2Z}{n^2h^2} \tag{36.7}$$

or

$$\boxed{E_n = -\frac{2Z^2\pi^2e^4m}{h^2}\frac{1}{n^2}} \tag{36.7a}$$

(where $n = 1, 2, 3, 4, \ldots$). Equation (36.7a) determines **the energy levels of the hydrogen atom.** *The only variable in this expression is the principal quantum number n, which is an integer.*

According to postulate (36.5a) the frequency of an emitted photon is given by the expression

$$\nu = \frac{E_0 - E_f}{h}$$

The two energies E_0 and E_f can differ only in the value of n in the energy expression (36.7a). We indicate this by writing E_n and E_k instead of E_0 and E_f, where k and n are different integers ($n > k$). Our **frequency equation** then becomes

$$\boxed{\nu = \frac{E_n - E_k}{h}} \tag{36.5b}$$

Introducing the wave number $\tilde{\nu}$ instead of the frequency ν ($\tilde{\nu} = 1/\lambda = \nu/c$, see Chap. 35) we can write for (36.5*b*)

$$\nu = c\tilde{\nu} = \frac{E_n - E_k}{h}$$

and hence,

$$\tilde{\nu} = \frac{1}{ch}(E_n - E_k) \tag{36.5c}$$

Substituting the values of E_k and E_n from equation (36.7*a*) and factoring out the constants, we get

$$\tilde{\nu} = -\frac{1}{ch}\frac{2Z^2\pi^2e^4m}{h^2}\left(\frac{1}{n^2} - \frac{1}{k^2}\right) \tag{36.8}$$

or

$$\boxed{\tilde{\nu} = \frac{2Z^2\pi^2e^4m}{ch^3}\left(\frac{1}{k^2} - \frac{1}{n^2}\right)} \tag{36.8a}$$

Comparing this equation with the series formula of Chap. 35,

$$\tilde{\nu} = R_H\left(\frac{1}{k^2} - \frac{1}{n^2}\right) \tag{35.2}$$

we realize that we have derived a correct law of spectral emission, at least as far as its general form is concerned. It remains to be seen if we actually get the correct value for the empirically determined Rydberg constant R_H by substituting the known values for the electronic mass and charge, for the velocity of light c, and for Planck's constant in the following expression, which is obtained from a comparison of equation (35.2) with equation (36.8*a*):

$$R_H = \frac{2\pi^2e^4m}{ch^3} \quad (Z = 1 \text{ for hydrogen}) \tag{35.2a}$$

Using the data known at that time Bohr obtained a value of R_H which agreed with the experimental value within 6 per cent, which was within the limit of experimental error. Using modern values of the universal constants we obtain an agreement within about 0.7 per cent ($R_H = 109{,}677.76 \text{ cm}^{-1}$).

Bohr's derivation of equation (36.8*a*) and his computation of the Rydberg constant can be considered as one of the greatest triumphs of theoretical physics. The great significance of Bohr's theory lies not so much in the fact that it provided a mechanical explanation of purely optical phenomena as in the recognition that the classical laws of mechanics and electrodynamics become inadequate in the domain of atomic dimensions. Bohr showed that they can be used for small-scale phenomena only in conjunction with certain restrictive amendments which are suggested by the quantum theory. The energy exchange between interacting bodies, which was assumed to be a continuous process in classical physics, was assumed by Bohr to be discontinuous. Bohr's theory of spectra confirmed the confidence of physicists in the quantum theory based on Planck's derivation of his law of "black-body" radiation and on Einstein's interpretation of the photoelectric effect. Bohr's atomic theory proved eminently fruitful in suggesting new laws and in interpreting previously unexplained phenomena.

Bohr's Picture of the Emission and Absorption of Light

Let us now examine more closely the process of emission and absorption of radiation according to Bohr's atomic model. Equation (36.6*a*) enables us to draw a diagram of the stable electron orbits for hydrogen (Fig. 36.2).[†] Equation (36.7*a*) makes it possible to compute the electron energies corresponding to these orbits. Arrows indicate the electron transitions which give rise to spectra of the different series. For $k = 1$ in equation (36.8*a*) we obtain lines of the Lyman series. The corresponding transitions are shown in Fig. 36.2 by arrows ending at the innermost orbit. Jumps down to the second orbit are shown to give rise to the Balmer series, and jumps down to the third orbit to the Paschen series ($k = 3$). But how do the electrons get into the higher orbits from which they can jump into orbits of lower energy? In a hot gas they may be raised from the ground orbit into a higher one by absorbing energy in a collision with a neighboring atom or molecule. *The atom is then said to be "excited."* Bombardment with fast charged

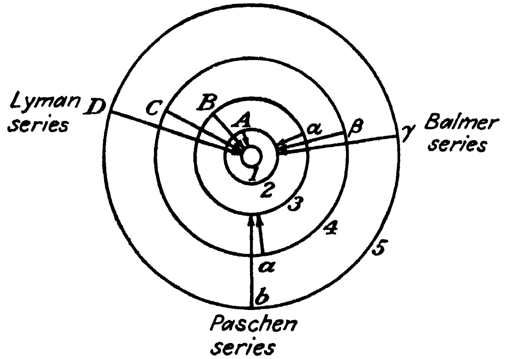


FIG. 36.2. Diagram illustrating emission of different spectral series by the hydrogen atom.

[†] The orbital radii are not drawn to scale. They should increase as n^2 according to equation (36.6*a*).

particles, such as ions or electrons, may also excite the atom. Excitation may be also produced by photons. When the gas is traversed by visible light of a continuous spectrum, only those photons are absorbed which possess precisely the energy required to raise an electron from the level in which it is encountered to a higher "permissible" energy level. In a cold gas we can assume practically all electrons to be in the "ground" orbit. Thus we shall expect photons corresponding to the frequencies of the Lyman series to be eliminated from the continuous spectrum by absorption, and we shall look for absorption lines in the same positions where the Lyman emission lines appear.† We shall, however, not expect to find absorption lines corresponding to the Balmer and Paschen series, since hardly any electrons will be present in the second and third orbits.

We can now understand the difference between the emission and absorption spectra. Most spectral series may be obtained in emission, since we can transfer the planetary electron by excitation to a high enough orbit (in fact, we can remove it by ionization), from which it has a choice of jumping back to the first, second, or third orbit, giving rise to lines of the Lyman, Balmer, or Paschen series. In absorption we can get only lines which correspond to the series emitted as a result of electron jumps into the "ground orbit." In hot gases, which may contain excited atoms, absorption lines of other series may appear.

We can now also understand Stokes' law of fluorescence; i.e., we can deduce it from Bohr's atomic theory. Whenever the absorption of a photon by an atom leads to the removal of an electron to a considerably higher orbit, the electron has a choice of returning to the original orbit in one jump or in cascade. For instance, suppose the electron of hydrogen has been raised from the ground orbit to the fourth orbit by absorbing a photon corresponding to the third line (C in Fig. 36.2) of the Lyman series. This electron can jump back to orbit 1 and give rise to the emission of a photon of the same wave length as had been absorbed. It can, however, also return to orbit 1 in cascade fashion, making the following alternative series of jumps: (a) $4 \rightarrow 3$, $3 \rightarrow 2$, $2 \rightarrow 1$ (three photons emitted); (b) $4 \rightarrow 3$, $3 \rightarrow 1$ (two photons emitted); (c) $4 \rightarrow 2$, $2 \rightarrow 1$ (two photons emitted). In all these cases the emitted photons are of smaller energy (and hence of lower frequency) than the originally absorbed photon (which corresponds to the largest jump), so that *the emitted fluorescent radiation must be either of the same or of longer wave length than the exciting radiation* (**Stokes' law**).

---

† This discussion is limited, for simplicity, to the consideration of dissociated monatomic hydrogen gas. The observations at room temperature and atmospheric pressure are complicated by the fact that two H atoms combine to form a molecule.

The Franck-Hertz Experiment as a Confirmation of Bohr's Theory

The actual existence of discrete atomic energy levels has been demonstrated convincingly by J. Franck and G. Hertz. They observed that *slow* electrons may suffer millions of collisions with molecules or atoms in a gas without a noticeable loss of kinetic energy (elastic collisions), but when their speed is increased to a certain minimum value, they lose all their kinetic energy, and at the same time sensitive optical devices reveal that this process of *inelastic collision of electrons with gas atoms* is accompanied by the emission of electromagnetic radiation (ultraviolet or visible). This observation can be interpreted as follows: As long as the kinetic energy of the bombarding electrons is smaller than the energy required to lift the planetary electron from the first to the second orbit, the atom is incapable of absorbing energy from the bombarding electrons. As soon as the electronic projectiles have acquired enough energy to excite the planetary electron to a transition to a higher orbit, the impact becomes inelastic. The atom absorbs the required *excitation energy*, thus robbing the bombarding electron of a part or of all of its kinetic energy.

Spectroscopic observation agrees with the predictions of the theory. As we increase the voltage which accelerates the bombarding electrons, we observe no emission of radiation until a certain minimum potential difference equal to V_1 has been reached. At this voltage the first line of the Lyman series appears. We conclude that V_1e is the energy required to lift the electron from the first to the second orbit. According to Bohr this energy must be equal to the energy of the photon which is emitted when the electron jumps back from the second to the first orbit. If our theory is correct, measurements must verify the following equation:

$$\boxed{V_1e = h\nu} \tag{36.9}$$

This prediction was actually confirmed by experiment. As we now increase the voltage, no new spectral lines are seen until a new accelerating voltage V_2 has been reached. Then two *additional* lines appear in the spectrum. They are the second line of the Lyman series and the first line of the Balmer series. This observation is interpreted by Bohr's theory as follows: At V_2 the bombarding electrons have acquired enough energy to transfer the planetary electron from the first to the third orbit. From there the electron may jump to the original orbit in two ways: (*a*) $3 \rightarrow 2$ followed by $2 \rightarrow 1$ (emission of two photons of frequencies: ν_{32} and ν_{21}) and (*b*) $3 \rightarrow 1$ (emitting light of frequency ν_{31}). The jump $3 \rightarrow 1$ leads to the emission of the second Lyman line, the jump $2 \rightarrow 1$ of the first Lyman line, and the jump $3 \rightarrow 2$ of the first Balmer line. The following energy relations are predicted by the theory.

For the direct jump to the ground orbit: $V_2e = h\nu_{31}$
For the cascade jump to the ground orbit:

$$V_2e = h\nu_{32} + h\nu_{21} = h(\nu_{32} + \nu_{21})$$

Comparing these two equations we see that the following relation must exist between the observed frequencies if the theory is correct:

$$\nu_{31} = \nu_{32} + \nu_{21}$$

This relation is in excellent agreement with observations.

As we continue increasing the accelerating voltage, we encounter more and more threshold values of V at which new spectral lines appear. We interpret the products V_ne (threshold voltage times the electronic charge) as the energy required to lift the planetary electron from the ground level to the nth level. Finally, we reach a voltage V_i at which the gas is ionized. *All* spectral series are now emitted by the atom (why?).

The success of verifying by collision experiments the predicted energy levels of the hydrogen atom and of other atoms can be considered a brilliant confirmation of Bohr's atomic theory.

The Structure of Heavier Atoms and the Periodic System of Elements

Bohr did not limit himself to the model of the hydrogen atom. He and Stoner proposed a scheme enabling one to construct models of heavier atoms. They started from the fact (mentioned in Chap. 35) that the nuclear charge of an atom is equal to its atomic number. Thus, for instance, an atom of element 23 has a positive nuclear charge 23 times as high as the charge of the electron and, being neutral, has 23 planetary electrons. Uranium, the heaviest element known to Bohr, was ascribed a nuclear charge of $Z = 92$ elementary units (being the ninety-second element in Moseley's sequence) and an equal number of planetary electrons. Where are all these planetary electrons located? Bohr and Stoner assumed that they are distributed over the various Bohr orbits, which for qualitative arguments can be computed roughly (ignoring the interaction between the electrons) for a nucleus of a given Z from equation (36.6*a*) by allowing n to assume the sequence of values $n = 1, 2, 3, 4, \ldots$. We thus obtain a series of *shells* occupied by electrons. Instead of quantum numbers 1, 2, 3, . . . it is customary to use letters $K, L, M, N, \ldots$ to designate these shells. For reasons which cannot be discussed in this text (*Pauli's exclusion principle*), *it was assumed that the number of electrons in a shell could not exceed* $2n^2$, where n is the principal quantum number of the corresponding Bohr orbit. Thus we obtain Table 36.1 correlating shell symbol, quantum number, and the number of electrons which "fills" the shell.

Table 36.1

Shell symbol	*K*	*L*	*M*	*N*	*O*
Quantum number	1	2	3	4	5
Number of electrons	2	8	18	32	50

The figures in the lowest line of this table acquire an interesting significance if they are examined in relation to the periodic system of elements. Mendeléeff arranged elements in the sequence of increasing atomic weights (in modern tables it is the sequence of atomic numbers) and wrote them in horizontal rows, as shown in Table 36.2, so that elements of similar chemical properties were arranged in vertical columns. Thus, for instance, we find hydrogen and the monovalent alkali metals Li, Na, K, Rb, Cs in the first column, the bivalent alkaline-earth metals in the second column, the nonmetallic halogens in the seventh column, the inert gases in the eighth column, etc.

It is a remarkable fact that chemical properties occur periodically among the elements, making such an arrangement possible. But the periodicity is not limited to chemical properties alone. Also physical properties, such as melting point, coefficient of thermal expansion, compressibility, and spectral characteristics of elements, were found to show a periodicity which coincided with the chemical periods of Mendeléeff. It is plausible to assume that this phenomenon must be due to periodically recurring features in the structure of atoms as we progress from hydrogen to heavier elements.

Let us now compare the figures in the bottom line of Table 36.1 with figures which we obtain by subtracting from each other the atomic numbers of two elements which follow each other in a vertical column of the periodic table. Let us consider the first column. We obtain (1) for H and Li, $3 - 1 = 2$; (2) for Na and Li, $11 - 3 = 8$; (3) for K and Na, $19 - 11 = 8$; (4) for Rb and K, $37 - 19 = 18$; (5) for Cs and Rb, $55 - 37 = 18$; (6) for Fa and Cs, $87 - 55 = 32$. We recognize in our results the figures for the number of electrons in the Bohr-Stoner shells given in Table 36.1. Similar results are obtained by examining other vertical columns (*e.g.*, column VII).

According to the Bohr-Stoner scheme, if we were to transform a hydrogen atom ($Z = 1$) into a lithium atom ($Z = 3$) by increasing the nuclear charge by two units and adding two electrons to preserve neutrality of the atom, the original electron of the hydrogen atom would be displaced to the *L* shell if the two new electrons are placed in the *K* shell, since this is the maximum number of electrons the *K* shell can accommodate. The lithium atom thus formed would resemble the a gap in the K shell. According to Bohr, electrons can jump now into the K shell from any of the upper shells. Since the energy of an electron is lower in the K shell than in the upper shells, this process accounts for emission of radiation. Since the "binding energy" [see equation (36.7*a*)] of the K electrons is very large for elements of high atomic number ($Z = 80$ for mercury), photons of very high energy (and hence high frequency) must be expected to result from jumps carrying an electron from, say, the P shell into the K shell. The estimated frequencies lie well in the X-ray region.

The knocking out of one K electron can lead to the emission of many photons of different frequencies. To mention only a few possibilities, (1) $P \rightarrow K$:† only one photon is emitted. A vacancy is created in the P shell. (2) $L \rightarrow K$: the first line of the K series is emitted. A vacancy is created in the L shell. Jumps of electrons into the vacancy in the L shell from any of the upper shells can now occur. They lead to emission of lines of the L series. Suppose, for instance, that the jump which occurs is $N \rightarrow L$, then the second line of the L series is emitted. We have now a vacancy in the N shell. Suppose it is filled by a jump $P \rightarrow N$, then a line of the N series is emitted. (3) $N \rightarrow K$ and $P \rightarrow N$. Two photons are emitted: (*a*) corresponding to the third line of the K series and (*b*) corresponding to the second line of the N series.

It is left to the student to give additional examples of such *cascade jumps*.

The above discussion shows how the mechanism proposed by Bohr can account in a qualitative way for the emission of **characteristic X rays,** *i.e.*, of a line spectrum which is typical for a particular element. Bohr also succeeded in offering a quantitative support for his theory of X-ray spectra. Moseley's law for the wave number of the K_α line of the X-ray spectra of various elements (*i.e.*, of the most intense line in the K series) can be roughly written in the following form according to empirical data:

$$\tilde{\nu} = R(Z-1)^2\left(\frac{1}{1^2} - \frac{1}{2^2}\right) \tag{36.10}$$

(R is very nearly equal to R_H, the Rydberg constant for hydrogen, and Z is the atomic number.‡)

This expression is very similar to the expression for the first line of the Lyman series of the hydrogen spectrum which is obtained from equation (36.8*a*) by setting $Z = 1$, $k = 1$, and $n = 2$. Comparing equation (36.10) with (36.8*a*), the latter equation being written for a hydrogenlike atom of atomic number Z emitting the first Lyman line,†

---

† This symbol means an electron jumps from the P shell to the K shell.

‡ Compare (36.10) with the square of equation (34.4) by setting $b = 1$ and $a^2 = R[(1/1^2) - (1/2^2)]$.

$$\tilde{\nu} = RZ^2\left(\frac{1}{1^2} - \frac{1}{2^2}\right) \tag{36.8b}$$

we notice that they differ only in that the factor Z^2 in equation (36.8*b*) is replaced by $(Z - 1)^2$ in equation (36.10). *The expression for the frequency of X rays emitted by various heavy elements is thus seen to be very similar to Bohr's expression for the frequencies of the lines of the hydrogen spectrum.*

Equation (36.10) holds for the emission of a K line. If we imagine a K electron knocked out of the atom, the atom will be ready for the emission of a line of the K series by an electron jump into the vacancy in the K shell. The single electron which has remained in the K shell, whirling about the nucleus of charge Z (units of electronic charge), can be imagined as "screening off" part of the nuclear charge, so that the electrons in the other shells which are ready to jump to a lower shell "see" a nuclear charge approximately equal to $(Z - 1)$ rather than a charge Z, which they would have seen in the case of a truly hydrogenlike atom, containing no electrons between the electron which is ready to jump and the nucleus. Thus we can apply Bohr's equation (36.8*b*) for the computation of the X-ray frequencies of the K series, merely replacing Z by $(Z - 1)$.‡

Whereas an atom can emit lines of any of the X-ray series when one of the K electrons has been removed, the X-ray absorption spectrum cannot be expected to be entirely identical with the emission spectrum. Since most of the shells of our model atom (Fig. 36.3) are fully occupied ("closed"), a photon which imparts to a K electron enough energy to lift it to, say, the L shell (which is filled) cannot be absorbed. Experience shows that only those X-ray photons are absorbed which supply to an electron at least the amount of energy sufficient to remove it beyond the region occupied by closed shells. We can picture roughly the process of absorption of X-ray photons as ionization in which internal electrons are ejected from the atom. As a result, when measuring the coefficient of absorption for X rays, as we increase the X-ray frequency continuously, we notice sudden sharp rises in the value of the absorption coefficient every time the frequency reaches a value where the photons possess enough energy to knock out electrons from an inner shell. The sharp rise is followed by a gradual drop of the absorption coefficient. Thus as we increase the frequency of X rays, we obtain a sharp increase in absorption when $h\nu$ becomes high enough to eject electrons from the *M* shell. The subsequent drop in μ is followed by a second sharp rise when the frequency becomes high enough for ejection of *L* electrons. A further sharp absorption peak occurs when ionization of the *K* electrons begins.

---

† R is slightly different from R_H for reasons which cannot be discussed here.

‡ This theory is only a rough approximation.

The Success of Bohr's Theory

It is impossible to enumerate here all the triumphs of Bohr's atomic theory. We can mention but a few important milestones such as the following:

1. Computation of the Rydberg constant from values of universal physical constants.
2. Explanation of the characteristics of series spectra and theoretical computation of known spectral frequencies for hydrogen and ionized helium.
3. Prediction of yet undiscovered spectral series (such as the Lyman series of hydrogen).
4. Explanation of periodicity in the chemical and physical properties of elements.
5. A more complete explanation of the Zeeman effect than was possible heretofore.
6. Explanation of differences between emission and absorption spectra.
7. Explanation of slight differences between the spectra of isotopes.
8. Explanation of the process of fluorescence and deduction of Stokes' law.
9. Theoretical derivation of Moseley's law and explanation of the characteristics of emission and absorption spectra of X rays.

The Failure of Bohr's Theory

However, after an initial period of enthusiasm physicists began to wonder if Bohr's theory was the ultimate word on atomic structure. (1) They were disturbed by the fact that no satisfactory theory could be devised for the emission of spectra by elements containing more than one electron in the outer shell. (2) The Zeeman effect could be explained only with strong magnetic fields, but the Zeeman splitting of spectral lines in weak fields could not be interpreted on the basis of Bohr's theory. (3) Finally, the following rather direct attempt was made to test the "correctness" of Bohr's model: Bohr's hydrogen atom is pictured as a disk (the plane of the disk being the orbit of the revolving electron). If it were possible to orient these atomic pancakes, should not the atoms oriented as shown in Fig. 36.4*a* encounter less resistance if shot through an atmosphere of hydrogen gas than would the atoms oriented as illustrated in Fig. 36.4*b*? But how could we orient the hydrogen atoms? If Bohr is right, the hydrogen atom should, owing to its revolving electron, behave in a magnetic field like a galvanometer coil. The plane of the orbit of the electron should orient itself at right angles to the magnetic field. *In Fig.* 36.4*b*, *the magnetic field is parallel to the velocity vector.* The disk-shaped atoms are oriented so as to offer maximum cross section for collisions with encountered gas atoms.

In Fig. 36.4*a*, *the magnetic field is applied at right angles to the plane of the paper* while the direction of motion is indicated by the $\vec{v}$ vector. In this case the disks offer a smaller cross section to surrounding gas atoms with which they collide, and so we expect the beam of hydrogen atoms shown in Fig. 36.4*a* to penetrate farther into an atmosphere of hydrogen gas than the beam shown in Fig. 36.4*b*.

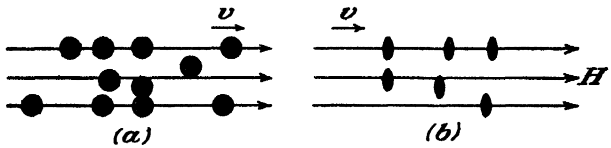


FIG. 36.4. Orientation of moving hydrogen atoms in a magnetic field. *a*. The magnetic field is perpendicular to the plane of the paper. *b*. The magnetic field is parallel to the direction of motion of the atoms.

Experiments showed that the magnetic field has no effect at all on the range of the atomic beam of hydrogen, suggesting that the hydrogen atom is spherically symmetrical rather than disk-shaped as pictured by Bohr.

The mounting evidence of the inadequacy of Bohr's atomic theory stimulated physicists to look for a new foundation of atomic mechanics. The new theory, which was finally evolved through the work of De Broglie, Schrödinger, Heisenberg, Dirac, and others, will be discussed in Chap. 37.

QUESTIONS AND PROBLEMS

1. What is the theoretical meaning of the "series limit" in terms of Bohr's model?

2. Is it conceivable for an atom to emit a continuous spectrum of frequencies exceeding the frequency of a series limit?

3. Are there (*a*) any atoms, (*b*) any ions whose spectrum should closely resemble the hydrogen spectrum? Give examples.

4. How should the spectrum of (*a*) singly ionized helium, (*b*) doubly ionized helium, (*c*) doubly ionized lithium differ from the hydrogen spectrum? Derive a series formula for each case and discuss it.

5. What happens when a hydrogen atom absorbs a light quantum whose wave length is shorter than that of the series limit of the series which is expressed by the following formula?

$$\tilde{\nu} = R_H\left(\frac{1}{1^2} - \frac{1}{n^2}\right) \qquad \text{(where } n > 1\text{)}$$

6. Compute the ionization energy of hydrogen, *i.e.*, the energy required to remove the planetary electron.

7. Compute the ionization energy for a singly ionized helium atom.

8. Does hydrogen absorb X rays? What do you expect to observe when X rays pass through an atmosphere of hydrogen?

9. Does hydrogen emit a characteristic X-ray spectrum? Does lithium emit one? Why or why not?

10. In what way does the X-ray emission spectrum differ from the X-ray absorption spectrum?

11. Compute the radius of the second Bohr orbit for a singly ionized helium atom.

12. Estimate as closely as you can the wave length of the K_α line of the X-ray spectrum of (*a*) uranium, (*b*) iron.

13. Study the periodic table with reference to the valence of elements, and formulate a hypothesis correlating the electronic structure of the atomic outer shell with the valence.

★**14.** Study chemical reactions between elements such as the alkali metals and the halogens, the alkaline-earth metals and oxygen or sulfur; consider the lack of chemical reactivity of the inert gases; and attempt to formulate a hypothesis which will answer the following questions:

a. Why do inert gases not combine with other elements?

b. Why are halogens monovalent toward metals?

c. Why can phosphorus display more than one valence?

d. Why is the valence of an alkaline-earth metal equal to the number of electrons in its outer shell?

e. Why do sodium ions in solution not react with water?

f. What forces hold Na and Cl together in an NaCl molecule?

15. Find approximately the maximum speed of the photoelectrons ejected from aluminum by the first line of the *K* series of iron.

16. Find the potential difference through which electrons must be accelerated in order to be able to ionize a hydrogen atom by collision.

17. Find the minimum speed of electrons with which one must bombard hydrogen atoms in order to excite the emission of the second line of the Balmer series.

18. Compute the speed of the electron in the first Bohr orbit of hydrogen.

19. Compute the radius of the first Bohr orbit for (*a*) hydrogen, (*b*) doubly ionized lithium, (*c*) triply ionized beryllium.

20. Compute the excitation energy required to transfer the electron of the hydrogen atom from the ground state to the second orbit.

★**21.** Develop a theory of the "positronium" atom, *i.e.*, an atom which so far has not been discovered and which is imagined to be similar to the hydrogen atom, except that a positive electron takes the place of the proton. Derive an expression for the energy levels of this atom.

CHAPTER 37

WAVE MECHANICS

a. Name the experiment which you consider the most convincing evidence in support of the photon theory of light.

b. In what way does the Compton effect differ from the photoelectric effect?

c. Under what conditions does light display (1) particle characteristics? (2) wave characteristics?

d. How can you predict in what direction a slow elementary particle will move after passing through an exceedingly small hole?

e. What reliable predictions can be made about the direction of motion of very small particles which are passing through a tiny hole?

f. What is the meaning of the equation $\lambda = h/mv$?

g. What is the ratio of the de Broglie wave lengths of an electron and a hydrogen nucleus moving at equal speeds?

h. What replaces Newton's first law in wave mechanics?

i. State the uncertainty principle.

j. What is the physical meaning of the amplitude of matter waves?

k. What attitude does wave mechanics take toward atomic models?

l. How does Schrödinger picture the electronic environment of a nucleus?

m. What is a potential barrier?

n. In what way is the "tunnel effect" inconsistent with classical mechanics?

o. In what way is the statement justified that the laws of quantum mechanics are statistical in nature?

p. What is the attitude of modern physics toward the principle of causality?

The conclusion of Chap. 36 must have been very disappointing to the student. After the demonstration of the brilliant initial success of Bohr's atomic theory, he was confronted with serious discrepancies between experiment and the predictions of that theory and was led to recognize its insufficiency. What can be wrong with Bohr's theory? Could it be that Coulomb's law, which had been used by Bohr, had been extrapolated to excessively small distances (of the order of magnitude of 10^{-8} cm)? This cannot be the reason, since Rutherford's alpha scattering experiments have demonstrated that Coulomb's law still holds well for interactions between "point charges" at distances of about 10^{-12} cm! The application of Planck's and Einstein's quantum-theoretical ideas to the atom could hardly be made responsible for the failure either. In fact, it was precisely the use of these ideas that enabled Bohr to obtain a theory leading to a good quantitative account of many phenomena which classical theory could not explain even qualitatively. Thus the only remaining thought is the rather disturbing assumption that the basic Newtonian laws of mechanics, which were used by Bohr, do not apply to subatomic particles moving in the submicroscopic domain of an atom. Bohr's theory was founded on the laws of Newtonian mechanics which, in order to lead to correct predictions, had to be supplemented by certain artificial quantum-theoretical assumptions. The use of quantum theory was undoubtedly a step in the right direction. Could it be that a more perfect agreement between theory and experiment would result if the classical mechanics, patched up by quantum postulates, were to be replaced by an entirely new system of mechanics purged of prejudices established by our experience with large-scale phenomena and especially adapted to yield correct prediction of atomic processes? The answer to this question was in the affirmative. A new, eminently successful system of mechanics, different from Newtonian mechanics, eventually evolved in the form of **wave mechanics.** Its basic idea was suggested by a comparison between the laws of mechanics and optics.

Mechanical Explanation of an Optical Phenomenon†

We shall begin our discussion of the "mechanics of light" and "optics of matter" by considering an experiment which can be regarded as the most direct demonstration of the particle nature of light. Compton was studying scattering of monochromatic X rays. One would expect the spectrum of the scattered X radiation to consist only of one single spectral line identical in frequency with the radiation which impinges upon the scattering body. This is a necessary consequence of a theory which assumes the scattered radiation to be reradiated by oscillating electrons which are set in vibration by the impinging wave.

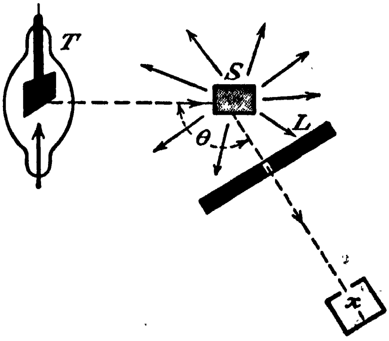


FIG. 37.1. Arrangement for the observation of the Compton effect.

Figure 37.1 shows schematically Compton's setup. X rays (dotted line) coming from the tube T strike the scattering body S, from which scattered X rays issue in all directions. X rays scattered at an angle to the direction of the original beam are allowed to pass through a hole in a lead diaphragm L and enter an X-ray spectrograph x. Compton did find, as expected, the characteristic line present in the original beam in the spectrum of the scattered radiation, *but in addition to it, he found a much weaker line of slightly longer wave length* which was absent in the spectrum of the original beam. The wave length of this "Compton line" was independent of the scattering substance, but it was found to depend on the angle of scattering θ. The Compton line was farthest removed from the expected line in the case of back scattering ($\theta = 0$), and its deviation from the normal line went toward zero as θ approached 180°.

---

†The word "optical" is used here in a wider sense, meaning electromagnetic radiation.

From the standpoint of the wave theory, it is impossible to understand how a radiation different in frequency from the incident beam could be emitted by the scatterer, but a computation in which X rays were considered as a hail of photons which were treated like elastic balls enabled Compton to explain his observations qualitatively, as well as to derive a quantitative expression for the separation $\Delta\lambda$ of the two spectral lines in the scattered radiation as a function of the scattering angle θ.

We shall treat this problem only qualitatively. Let us visualize a photon as an elastic particle of energy $E = h\nu$. According to Einstein's law (10.7), we can ascribe to it a mass of

$$\boxed{m = \frac{E}{c^2} = \frac{h\nu}{c^2}} \qquad (37.1)$$

Since the velocity of the photon is c, its momentum is given by

$$p = mc = \frac{h\nu}{c^2}c = \frac{h\nu}{c} \qquad (37.2)$$

This equation can be written in terms of the wave length λ, making use of the relation $c = \nu\lambda$. Substituting $\nu/c = 1/\lambda$ in equation (37.2), we get

$$\boxed{p = \frac{h}{\lambda}} \qquad (37.2a)$$

We can now explain the result of Compton's scattering experiment by the following mechanical consideration.

1. **Normal Scattering** (λ remains unchanged). Some of the photons penetrate deeply into the atom and collide with one of the inner shell electrons which are firmly bound to the nucleus, being very close to it. The inner electron can be considered to be quasi-rigidly attached to the nucleus, so that the collision is equivalent to the collision of the photon with the entire massive atom. It bounces off as if from an elastic wall of infinite mass, retaining practically all its original momentum and energy. This explains the presence of the line of unchanged wave length in the scattered radiation, since the constancy of p implies the constancy of λ according to equation (37.2a).

2. **Compton Scattering** (λ is lengthened). Some of the photons collide with one of the outer electrons which is far from the nucleus and is so

loosely bound that it can be practically considered as a free electron. The mass of the photon is comparable to the mass of the electron, and the collision resembles the collision between two elastic balls. The electron acquires momentum, and hence energy, through the impact at the expense of the photon. It is easy to see from equation (37.2*a*) that, since h = constant, *the loss of momentum of the photon must be accompanied by an increase in* λ. This explains the presence of the second line of lengthened wave length in the scattered radiation observed by Compton. A simple calculation based on the laws of conservation of energy and of momentum leads to an expression relating the change in λ to the scattering angle θ. The excellent agreement between this deduced formula and experimental observation can be considered a very convincing confirmation of Einstein's photon hypothesis.

An Optical Phenomenon Not Explicable on the Basis of a Mechanical Particle Theory

Encouraged by Compton's success in applying the law of conservation of momentum (derivable from Newton's second and third laws) to the interaction between photons and electrons, we shall now proceed to check the applicability of Newton's first law to the motion of photons. According to the law of inertia, the photons of a parallel beam of light L (Fig. 37.2) which pass through the slit S in the diaphragm D should, if they are corpuscles obeying Newtonian mechanics, continue on a straight line toward the screen Sc, where they should illuminate an area identical in size and shape with the slit S.

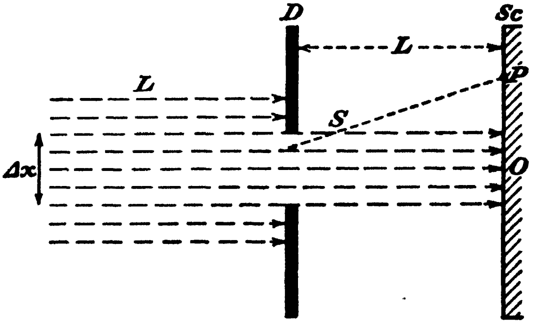


FIG. 37.2.

As we know, this prediction holds quite well only for slit openings which are very large compared with the wave length of light. But when the slit width Δx is comparable to λ, the illuminated area of the screen Sc is considerably wider than Δx and is traversed by dark interference fringes (Fig. 29.9). Figure 29.10 shows how the distance x_1 of the first interference minimum from the center point O can be computed. In the present case this distance is given by equation (29.5) when we replace D by Δx:

$$x_1 = \frac{\lambda}{\Delta x} L \tag{37.3}$$

By a similar consideration the distances of the second, third, and other minima are shown to be

$$x_2 = \frac{2\lambda}{\Delta x} L \tag{37.3a}$$

and

$$x_3 = \frac{3\lambda}{\Delta x} L \text{ etc.}$$

The fact that light is detected at points within the area of the geometrical shadow, such as P of Fig. 37.2, shows that photons do not obey the law of inertia. One might think that we have no reason to doubt the validity of the law of inertia in this case, since the deviation of photons from the linear path could be explained by forces of interaction between neighboring photons. But diffraction experiments have been performed at such low light intensities that at any one moment only one photon could have been present in the space between the slit and the screen. The usual interference patterns were detected at these low intensities of radiation. Under such conditions interaction between photons is ruled out, and we conclude that *the motion of photons past narrow confines cannot be accounted for by the laws of Newtonian mechanics.*

The distribution of the intensity of illumination in the interference pattern (Fig. 29.9) is determined by the variation of the amplitude of the stationary light waves formed behind the slit. The intensity of illumination is measured by the square of the amplitude of the electric or magnetic field vector of the light wave. We have learned in Chap. 35 how the square of this amplitude at a point has been correlated by Einstein with the photon density, which determines the rate of incidence of the photons per unit area at that point. Thus *the interference pattern appearing on the screen is described as follows in terms of the photon theory:* Most of the photons are streaming toward the central bright portion opposite the slit; the stream of incident photons diminishes toward the first dark fringes where almost no photons are landing. From there on the photon stream increases again toward the first lateral maxima and drops to nearly zero at the second dark fringes, etc. The photons move as if they were guided by the electromagnetic waves rather than following the law of inertia.

The Uncertainty Principle for Photons

When a photon passes through the slit in the above experiment we have no way of anticipating whether it will go toward the first, the second, or the third bright fringe and where exactly it will strike the screen. When we attempt to define the location of the photon more sharply within the slit by narrowing its width, we increase the indefiniteness of the direction in which the photon is likely to proceed after passing the slit. A glance at equations (37.3) and (37.3*a*) reveals that, as we diminish the width Δx of the slit, the possible deviation of the photons from their original course increases; *i.e.*, the dark fringes, whose location is a measure of the maximum sideways deviation of the photons, shift farther away from the center. (x_1, x_2, etc., become larger.) Thus while we gain a better definition of the location of the photon within the slit, we cause the photons to be scattered through a wider angle.

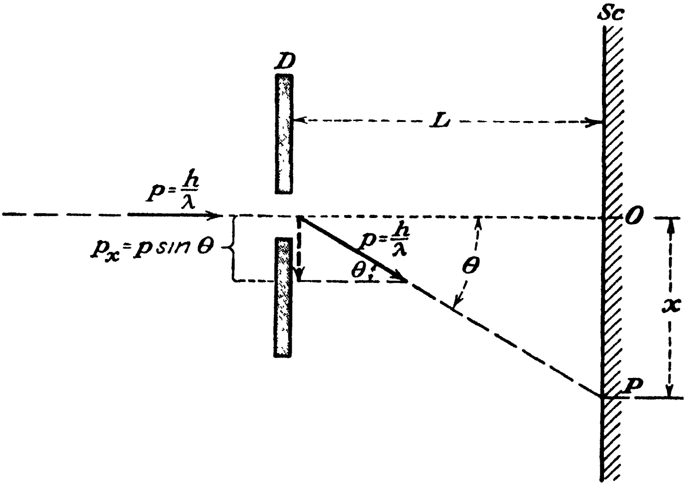


Fig. 37.3.

These conditions can be formulated in an approximate fashion as follows: In Fig. 37.3 the horizontal broken line indicates the path of a photon which passes somewhere through the slit in the diaphragm *D*. Let us assume that, after passing through the slit, the photon continues moving with the same momentum† in a different direction, say toward a point *P*, which is located between the two first dark fringes. The photon has obviously acquired a sideways component of momentum which can be read off from Fig. 37.3.

$$p_x = p \sin \theta = \frac{h}{\lambda} \sin \theta \tag{37.4}$$

---

† We know that $p = h/\lambda$ has not changed in magnitude, since λ remains the same.

For small values of θ, we can write

$$\sin\theta \simeq \tan\theta \simeq \frac{x}{L}$$

and hence,

$$p_x \simeq \frac{h}{\lambda}\frac{x}{L} \tag{37.4a}$$

If the photon hits the screen somewhere between the center of the diffraction pattern and the first dark fringe, the deviation x cannot be any larger than the coordinate x_1 of the first minimum [equation (37.3)]. Hence, the maximum value of the lateral momentum would be obtained in this case by substituting x_1 from (37.3) in (37.4a):

$$p_x = \frac{h}{\lambda L}x_1 = \frac{h}{\lambda L}\frac{\lambda}{\Delta x}L = \frac{h}{\Delta x} \tag{37.5}$$

And in case the photon strikes the screen somewhere between the first and the second dark fringe, its deviation could not exceed x_2, the coordinate of the second dark fringe [equation (37.3a)]. Hence, in this case the maximum acquired lateral momentum would be obtained by substituting x_2 from (37.3a) in (37.4a):

$$p_x = \frac{h}{\lambda L}x_2 = \frac{h}{\lambda L}\frac{2\lambda}{\Delta x}L = \frac{2h}{\Delta x} \tag{37.5a}$$

and similarly we should find for photons striking within the adjacent bright zone (*i.e.*, between the second and the third dark fringes) the maximum lateral momentum

$$p_x = \frac{3h}{\Delta x} \qquad \text{etc.} \tag{37.5b}$$

Let us consider first the photons which strike the screen between the center and the first dark fringe. They may acquire a maximum lateral momentum of p_x toward either side of the slit (the two directions are distinguished by change of sign, so that we write p_x and $-p_x$). We can express this by saying that *their lateral momentum is indefinite to the extent* $\Delta p_x = p_x - (-p_x) = 2p_x$. Substituting the value of p_x from equation (37.5) in the above expression, we get

$$\Delta p_x = 2\frac{h}{\Delta x}$$

or

$$\Delta p_x\,\Delta x = 2h \tag{37.6}$$

Had we used the values of p_x as expressed by equations (37.5a) and (37.5b), which determine the indefiniteness of momentum for photons striking the screen between the first and second dark fringes and between the second and third dark fringes, respectively, we should have obtained

$$\Delta p'_x = 2\frac{2h}{\Delta x} \qquad \therefore \Delta p_x \, \Delta x = 4h \tag{37.6a}$$

$$\Delta p''_x = 2\frac{3h}{\Delta x} \qquad \therefore \Delta p_x \, \Delta x = 6h \tag{37.6b}$$

The expression $\Delta p_x \Delta x$ *can be considered the product of the indefiniteness of lateral momentum* (Δp_x) *multiplied by the indefiniteness in the position* (Δx) *of the photon.* From the equations (37.6) to (37.6b), we see that this product is of the order of magnitude of Planck's constant h. It is customary to express the above relations by one single approximate equation:[†]

$$\boxed{\Delta p_x \, \Delta x \gtrsim h} \tag{37.7}$$

This is a statement of the famous **principle of indefiniteness** (*or, as it is also called, principle of* **indeterminacy** *or of* **uncertainty**) *of Heisenberg.* **In its present application to photons it states that it is impossible to determine simultaneously with unlimited precision the location as well as the momentum of a photon.** If we decide to localize the photon sharply by reducing the slit aperture Δx, we increase the indefiniteness of its lateral momentum $\Delta p_x \gtrsim h/\Delta x$. On the other hand, we can reduce the indefiniteness Δp_x of its momentum by increasing Δx to any desired width, but then we increase the indefiniteness in the location of the photon at the moment when it passes the slit, since the photon is equally likely to pass at any point through the slit.

This principle is of more general importance than is apparent from the above discussion of its validity for photons. We shall review its general implications further on in this chapter.

Matter Waves

After the convincing demonstrations of the wave nature of light and of the existence of photons, we may ask ourselves in our deep confusion whether we have now to picture light as a wave motion or as a motion of particles. It has been suggested facetiously that light is due to a motion of "wavicles." We may, however, take this suggestion seriously. There is no reason why photons should resemble closely any object of our common experience such as, say, a sand particle or a familiar process such as a water wave. *We shall consider photons as entirely new elements of our experience which under certain conditions behave like particles and under others exhibit wave characteristics.* The latter behavior is observed when photons are made to pass through narrow confines, such as slits or small holes.

[†] The sign $\gtrsim$ means "approximately equal to and no less than."

What right do we have to assume that an electron or a hydrogen ion resembles a grain of sand any more than a photon does? Perhaps this dual behavior of partial resemblance to a wave and partial similarity to a particle is a general property of all elementary quanta of matter and energy. It may even be a general property of *all* moving matter without our being aware of it, because the manifestations of this dualism become distinctly noticeable only for very small bodies moving through very narrow confines. This is a peculiar thought, but let us toy with it for a little while. Modern physics abounds in examples of queer ideas which met with remarkable success.

If, for instance, electrons are associated with a wave process just as photons are, how could we become aware of it? We should expect to find a diffraction pattern similar to the one obtained with light (see Fig. 29.9) by passing electrons through narrow slits or diffraction gratings. We should have to abandon laws of Newtonian mechanics in predicting their motion and should have to make use of laws of wave optics to determine where the electrons will strike the screen *Sc*. *As in the case of photons (see above) we should consider the waves associated with the moving electrons as "guiding waves" determining by their amplitude the frequency of incidence of the electrons at various points of the screen.* We shall not make any other hypothesis about the nature of such **matter waves** except that *the square of their amplitude at a point is to determine the probability of finding a particle at that point.*

It is obvious that wavelike characteristics of electrons, which would make them behave according to laws different from those of Newtonian mechanics when passing through narrow confines (such as slits), would be of decisive importance for atomic theory, since the atomic electrons are confined to move in a tiny space whose order of magnitude is indicated by the atomic radius (about 10^{-8} cm). The modification of the laws of electronic motion required by our hypothesis may well turn out to be the new foundation of atomic mechanics for which we are searching.

As a final step in the formulation of our hypothesis, let us find a *way of assigning a wave length to a moving particle.* We shall link tentatively the momentum $p = mv$ of a particle with a wave length through the same relationship [equation (37.2*a*)] which holds for photons:

$$p = mv = \frac{h}{\lambda}$$

hence,

$$\boxed{\lambda = \frac{h}{mv}} \tag{37.8}$$

This is the famous **equation of de Broglie**, which he proposed in his Ph.D. dissertation as a reasonable hypothesis.†

Experimental Demonstration of Matter Waves. How can we test our hypothesis experimentally? The simplest method would be to send a parallel beam of electrons of velocity v, which we imagine to be associated with the matter wave of $\lambda = h/mv$, through a slit in a setup such as shown in Fig. 37.2 and to search for "dark" and "bright" fringes on the photographic plate (Sc), *i e.*, for zones of scarcity and of abundance of incident electrons forming an interference pattern similar to the illumination pattern shown in Fig. 29.9. In order to affect the photographic plate, the electrons should not be too slow. Let us say we use electrons accelerated through a potential difference of 54 volts. We can easily compute the velocity and hence the de Broglie wave length of these electrons. The velocity follows from the equation $\frac{1}{2}mv^2 = eV$, from which follows

$$v^2 = \frac{2eV}{m} = \frac{2(4.8) \times 10^{-10}(54)}{9.10 \times 10^{-28}(300)} = 19.0 \times 10^{16}\ (\text{cm/sec})^2$$

(where $\frac{54}{300}$ is the voltage in esu). Hence, $v = 4.36 \times 10^8$ cm/sec.

The wave length is, according to de Broglie,

$$\lambda = \frac{h}{mv} = \frac{6.62 \times 10^{-27}}{9.10 \times 10^{-28} \times 4.36 \times 10^8} = 1.67 \times 10^{-8}\ \text{cm} = 1.67\ \text{A}$$

This wave length is as short as that of an X-ray photon. We could not practically build a sufficiently narrow slit to demonstrate diffraction of such short waves. The problem is even greater with faster electrons or with particles of a larger mass. But could we not use the same diffraction methods as have been used to demonstrate the wave nature of X rays? It turns out that we can. Kikuchi sent electron beams (accelerated by a potential difference of about 70,000 volts) through very thin crystals of mica and got dot patterns similar to the diffraction patterns obtained by Laue with X rays (Fig. 37.4; compare Fig. 34.12). G. P. Thomson sent electron beams of 60,000 volts through thin gold foils containing microcrystals of gold. The experiment is analogous to the X-ray diffraction method of Debye and Scherrer illustrated in Fig. 34.21. Owing to the presence in the foil of tiny crystals oriented in all conceivable directions, a ring-shaped diffraction pattern (Fig. 37.5) was obtained using the setup of Fig. 37.6. Electrons accelerated in the electron gun emerge as a narrow beam and pass through the foil F, proceeding from there in certain preferential directions so as to give rise to a ring pattern on the photographic plate P of Fig. 37.6. *This ring pattern is interpreted as an interference pattern of the matter waves* which are associated with the beam of electrons. These waves are reflected by the microcrystalline structure only in certain preferred directions, on account of interference, in the same fashion as the X rays in the setup of Fig. 34.21. Electron diffraction was demonstrated prior to Kikuchi and Thomson by Davisson and Germer in a study of selective reflection of electrons from crystal surfaces. Just like X rays passing through a crystal, electrons are reflected from surfaces of crystals in

---

† In equation (37.8), v is the velocity of the moving particle. It is not to be confused with the velocity of propagation of the associated matter wave.

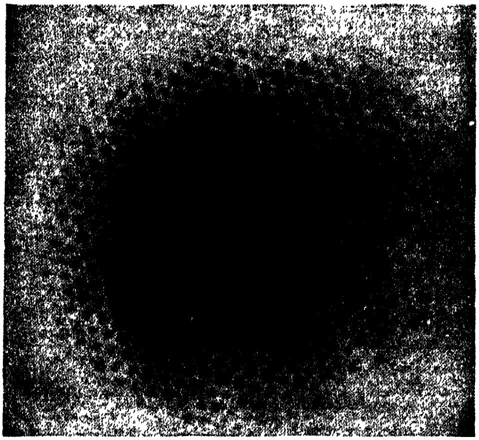


FIG. 37.4. Diffraction pattern obtained by passing 68,000-volt electrons through a thin sheet of mica. (*After Kikuchi.*)

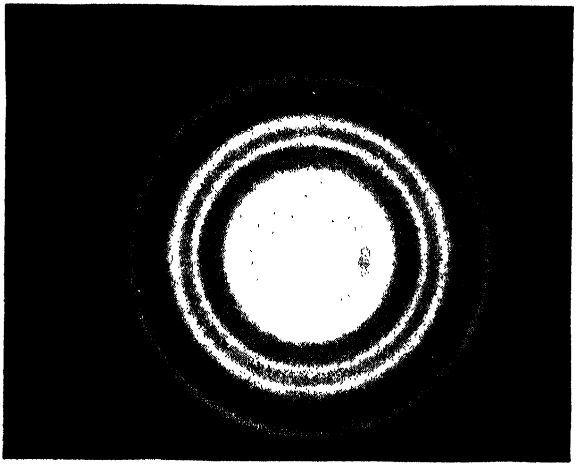


FIG. 37.5. Diffraction pattern obtained by passing 60,000-volt electrons through a thin gold foil. (*Courtesy of G. P. Thomson.*)

certain preferential directions owing to interference of matter waves reflected from adjacent layers of atoms. The angles of preferential reflection depend on the arrangement and separation of the atoms in the crystal lattice as well as on the de Broglie wave length of the electrons. *Davisson and Germer were able to determine λ from the diffraction pattern, i.e., from the direction of preferential reflection. Their measurements confirmed de Broglie's equation* (37.8) *for the first time. Later, Stern and coworkers succeeded in demonstrating diffraction of beams of atoms which were reflected from crystal surfaces.* **It was finally established by experiments that diffraction phenomena similar to those obtained by Davisson and Germer and by G. P. Thomson can be obtained not only with electrons but with charged or uncharged particles of any kind.** From this we infer that one could assign a de Broglie wave length even to a moving bullet. Why then don't we observe diffraction of bullets emerging from a machine-gun muzzle similar to the diffraction of photons emerging from a narrow slit? Compute λ for an average-sized bullet and you will have the answer!

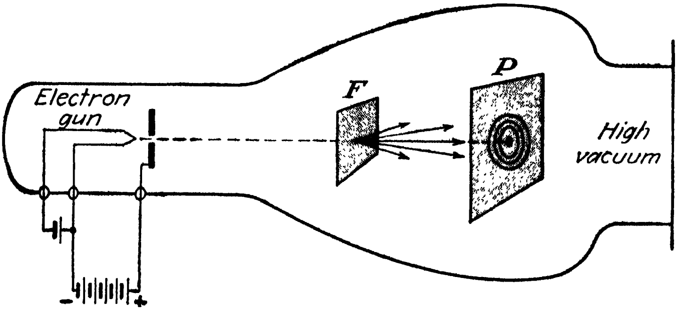


FIG. 37.6. Scheme of an arrangement for observation of electron diffraction by thin foils.

De Broglie Waves in an Atom

The confirmation of de Broglie's hypothesis by Davisson and Germer suggests that the motion of an electron inside an atom might be governed by its wave properties in a similar fashion to the motion of electrons between the atoms of a crystal lattice. Let us consider the motion of an electron in a circular orbit about the nucleus. We shall have to imagine the associated de Broglie wave to ring the orbit of the electron. An electron moving in a straight line with a definite momentum has a de Broglie wave of $\lambda = h/mv$ associated with it which extends to infinity on either side of the electron (Fig. 37.7).† In the case of motion in the circular orbit we imagine the infinite de Broglie wave wrapped around the orbit an infinite number of times. If the circumference of the orbit happens to be an integral multiple of the wave length

---

† The velocity u of the de Broglie wave is larger than the particle speed v. The electron e cannot be localized exactly at a definite point within the wave. It is nevertheless shown in Fig. 37.7 for clarity of the illustration.

$$2\pi r = n\lambda \tag{37.9}$$

the crests and troughs of the individual turns will coincide with each other as shown in Fig. 37.8*a*. If, on the other hand, $2\pi r \neq n\lambda$, the crests and troughs of the individual turns of the "wave coil" will be out of phase. Figure 37.8*b* shows the case where two consecutive turns are in phase opposition. Under these conditions pairs of consecutive wave turns cancel each other by interference, so that no de Broglie waves can persist along such orbits. *If we assume that a zero amplitude of de Broglie waves indicates the absence of electrons, we shall conclude that no electrons can move in such an orbit.* In the general case of an arbitrary orbital radius two consecutive "wave turns" will be somewhat out of phase and the phase difference between the first turn and the *n*th turn will increase with the number *n*. For every turn we could find an *n*th turn such that the two "wave turns" will be most nearly in phase opposition. Thus we see that the de Broglie waves cancel each other by interference more or less completely in all circular paths except those whose circumference is an integral multiple of λ. *This consideration leads us to the prediction that electrons could move only in orbits whose circumference is an integral multiple of* λ, *as expressed by equation* (37.9).

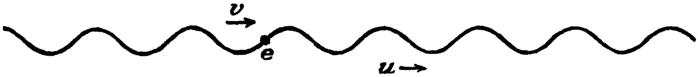


FIG. 37.7.

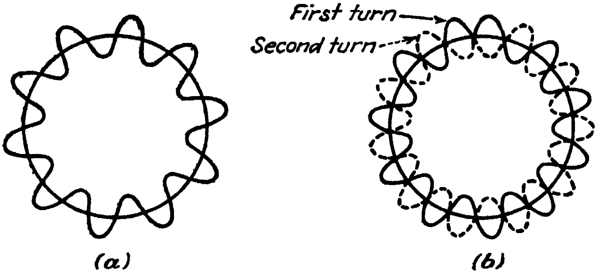


FIG. 37.8.

Substituting the value of λ from equation (37.8) in equation (37.9), we obtain

$$2\pi r = n\frac{h}{mv} \tag{37.10}$$

or

$$mvr = n\frac{h}{2\pi} \tag{37.10a}$$

But mvr is the angular momentum of the planetary electron [see (36.4*a*)]. We recognize thus that equation (37.10*a*) is identical with Bohr's postulate [equation (36.4)] stating that the angular momentum of a revolving electron can assume only values which are integral multiples of $h/2\pi$. *Thus an assumption which was made by Bohr arbitrarily and was justified only by the correct predictions to which it led appears as a natural consequence of de Broglie's hypothesis.*

Again we have to disillusion the student, who may have concluded that we have at last arrived at the "final" explanation of atomic phenomena. The picture of electronic orbits inside an atom has been rejected by modern physical theories, so that the above-mentioned elegant derivation of Bohr's quantization condition for the angular momentum of planetary electrons is mainly of historical interest.

The General Uncertainty Principle and Complementarity

The rejection of Bohr's planetary model is demanded by Heisenberg's uncertainty principle, which can be applied by the same argument to material particles as to photons. The implications of equation (37.7) have been generalized by Heisenberg as follows: *It is impossible* in principle *to determine simultaneously with unlimited precision both the location and momentum of a particle* (**uncertainty principle**).

In the case of photons passing through a slit it was seen that, owing to diffraction of the associated waves, a greater and greater lateral momentum was acquired (*i.e.*, the angular spread of the particles emerging from the slit was increasing) as we attempted to localize the position of the particles within the slit more sharply by diminishing its width Δx. The uncertainty of momentum was shown to be inversely proportional to the indefiniteness of position:

$$\Delta p_x \gtrsim \frac{h}{\Delta x} \tag{37.7a}$$

The same diagrams and considerations can be applied to moving material particles and the associated matter waves, leading to the same uncertainty relations [*equation* (37.7)]. Equation (37.7*a*) implies in relation to electrons that an attempt to reduce the vagueness Δx in the position of an electron to zero should make the uncertainty Δp_x in its momentum infinitely large. And conversely, an electron of perfectly defined momentum

($\Delta p_x = 0$) could not be localized anywhere in space, Δx being infinite in this case.[†] *Thus in ordinary considerations of an electron in motion we should make a compromise by admitting a reasonable indefiniteness in its momentum and in its position.*

This mutual exclusion of definiteness of momentum and of position has a deeper significance; it indicates that an observed phenomenon cannot be described at the same time in terms of the corpuscular theory and in the language of the wave theory. This can be seen as follows. According to the de Broglie relation [equation (37.8)] a definite wave length is assigned to a definite momentum. If we now describe a phenomenon in terms of the corpuscular picture and localize the particle sharply (*i.e.*, make $\Delta x = 0$), the momentum and hence also the de Broglie wave length become completely indefinite and we cannot describe the process in question in terms of the wave theory. On the other hand, if we define the wave length sharply ($\Delta\lambda = 0$ and hence $\Delta p = 0$), the position becomes completely indefinite [$\Delta x = \infty$ according to (37.7)] and we cannot use the description in terms of the particle theory.

According to Bohr, the wave aspect and the particle aspect of a process are two sides of reality which cannot be observed at the same time in a given process; either view alone is insufficient for the description of nature. There is a relation of **complementarity** *between the two modes of describing nature; i.e., the wave theory and the particle theory complement each other in the description of our experience.*

Thus, for instance, in computing the distribution of light intensity or of the intensity of a corpuscular radiation passing through an appropriate diffraction grating, we make use of the wave theory without having to consider the corpuscular aspect of these radiations. On the other hand, when we try to account for phenomena such as photoelectric effect, Compton effect, or deviation of moving electrons in a magnetic field, we make use of the corpuscular aspect, there being no indication of the wave nature of light or of the electrons in these experiments. *But we cannot interpret all observations made with light or material particles by restricting ourselves to only one of the two theories. Both the wave theory and the corpuscular theory are necessary for the interpretation of all the phenomena we observe with light and with electrons.*

[†] It is easy to see that indeterminacy of position and velocity is of importance only for particles of small mass. Let us write for the indefiniteness of momentum in equation (37.7*a*) $\Delta p_x = m\,\Delta v_x \gtrsim h/\Delta x$. From this follows

$$\Delta x\,\Delta v_x \gtrsim \frac{h}{m} \tag{37.7b}$$

We see that as the mass m increases, h/m goes toward zero and so does the product of the two indefinitenesses Δx and Δv_x. Thus simultaneous determination of position and velocity can be made with very high precision for sufficiently large bodies.

Abandonment of Mechanical Atomic Models

When we speak of an orbit of an electron, we imply that the electron can be found at any one moment at a point of its path line having a definite velocity. Such a definite picture is operationally meaningless according to Heisenberg's uncertainty principle. Suppose that in the description of a circular orbit of radius $r = 10^{-8}$ cm we want to localize the electron with an error of no more than $\Delta x = 10^{-11}$ cm, *i.e.*, within 0.1 per cent. How large will be the indefiniteness of momentum? From equation (37.7*a*) follows

$$\Delta p_x \gtrsim \frac{h}{\Delta x} = \frac{6.6 \times 10^{-27} \text{ erg-sec}}{10^{-11} \text{ cm}} = 6.6 \times 10^{-16} \text{ dyne-sec}$$

How indefinite does this make the velocity of the electron?†

$$\Delta p_x = \Delta(mv)_x = m\,\Delta v_x = 6.6 \times 10^{-16} \text{ dyne-sec}$$

$$\text{hence, } \Delta v_x = \frac{6.6 \times 10^{-16}}{m} = \frac{6.6 \times 10^{-16} \text{ dyne-sec}}{0.91 \times 10^{-27} \text{ gm}} = 7.4 \times 10^{11} \text{ cm/sec}$$

The uncertainty Δv_x comes out much greater than the velocity of light!

The uncertainty principle put an end to the search for the "correct" mechanical model of the atom. It marks the end of an era centuries long during which the description in terms of a mechanical model was considered as the ultimate goal of physical explanation. This principle implies that such explanations are reasonable only for large-scale phenomena where the indefiniteness of definition remains insignificant because of the smallness of h but become untenable on the scale of atomic dimensions for which h becomes noticeably large as compared with the mechanical constants of the system such as its angular momentum.

The Schrödinger Atom

But if we are to abandon the picture of the orbital electron model, how should we imagine the appearance of an atom? How are the electrons distributed in its interior? The answer to these questions was given by Schrödinger, who showed that the distribution of electrons in the atom is determined by stationary three-dimensional de Broglie waves which can be imagined to exist in the interior of an atom. To understand Schrödinger's idea it will be helpful to consider a mechanical analogy. We have seen in Chap. 27 that a string is capable of vibrating in different modes (Fig. 27.13). Such stationary-wave patterns are formed when waves reflected to and fro between two boundaries of a medium interfere with each other. In the above-mentioned example the string acts as a one-dimensional medium. A more complicated standing-wave pattern can be produced in a two-dimensional medium such as a drum membrane. When cork dust is spread over such a vibrating membrane, it is shaken off the regions of agitation and collects in the so-called *nodal regions* (designated by N in the diagrams) where the membrane is at rest. Figures 37.9*a* and *c* show circular nodes (N, N', and N''). [Linear nodal regions (N_a, N_b, N_1, N_2, N_3) can also be seen in Figs. 37.9*b* and *c*]. These diagrams correspond to snapshots at a given instant; if the shaded areas move up, the white ones move down. Membranes are capable of vibrating in many different modes, and an even greater variety of complicated nodal patterns is obtained with three-dimensional vibrating objects such as, say, a vibrating elastic sphere in which stationary waves are set up. Schrödinger assumed the de Broglie waves to form three-dimensional stationary waves in the interior of an atom. He reasoned, on the basis of his theory, that complicated nodal surfaces, *i.e.*, regions of zero amplitude of matter waves, should exist inside an atom. As seen from Fig. 27.13, in a stationary wave the amplitude is zero at the nodes and increases toward the antinodes. According to the wave mechanical interpretation of the amplitude of matter waves, the square of the amplitude measures the probability of encountering an electron at a given point. Hence, the nodal regions in Schrödinger's atom must be interpreted as "forbidden zones" for the electrons. The antinodes are the regions where the electron dwells most of the time. *The only assumption we make about the motion of our atomic electron is that its probability density*† *at a point is proportional to the square of the matter-wave amplitude at that point.* The over-all effect of the variability of the dwelling time of an electron in a certain region is the same as if the negative electric charge were present inside the atom in the form of a cloud of variable density which reaches a minimum in the nodal zones and a maximum in the antinodal regions.

---

† This estimate is very crude. The correct computation involves the use of relativistic formulas. It leads, however, to a similar conclusion.

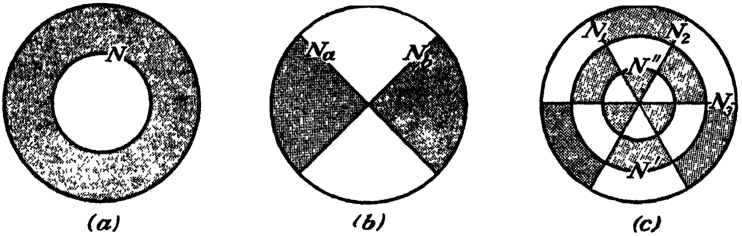


FIG. 37.9. Stationary-wave patterns for vibrating membranes.

---

† The probability density measures the likelihood of finding the electron in a unit volume.

Figure 37.10 shows a few diagrams of the many possible stationary-wave patterns for a hydrogen atom. The dark areas are nodal regions; they are zones where the probability of encountering an electron is negligible. The diagrams can be also interpreted as pictures of the density distribution of the negative cloud (shown white in the figure) surrounding the nucleus. They are to be visualized in space by imagining each diagram to be rotated about a vertical axis of symmetry. We see that, according to Schrödinger, the electronic charge of the hydrogen atom is distributed *spherically* about the nucleus in the ground state (Fig. 37.10*a*). This is in agreement with the experiments which failed to confirm Bohr's picture of a flat planetary hydrogen atom (see end of Chap. 36) and suggested that hydrogen atoms are spherically symmetrical.

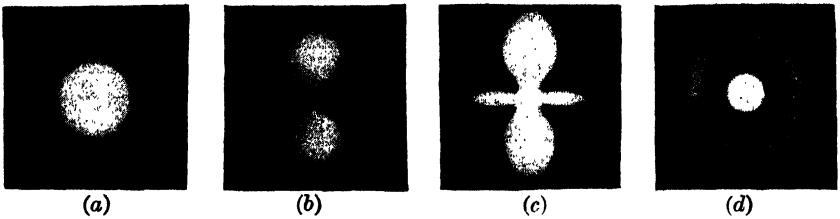


FIG. 37.10. Some of the possible distributions of the electron charge density for the hydrogen atom according to wave mechanics. *a*. The charge distribution for the ground state. *b*, *c*, and *d*. For higher excited states. (*From White, "Introduction to Atomic Spectra," McGraw-Hill.*)

The dark region between the inner white globe and the outer white shell of Fig. 37.10*d* corresponds to a nodal region. It is not a "forbidden zone" for the electron in so far as its presence in that space is not impossible but merely very improbable. *It is an important distinction between the modern quantum mechanical ideas about atomic structure and Bohr's theory that there are no sharply defined orbits or sharp boundaries of any kind determining the whereabouts of an electron.* An electron can be any place, up to infinity, according to modern theory, but certain regions exist in which its presence is more probable than in others. Wave mechanical theory shows that in heavier atoms there are zones of maximum probability of "electron residence" surrounding the nucleus in the fashion of spherical shells. The distances of these zones from the nucleus correspond closely to the radii of the K, L, M, and subsequent shells of the Bohr theory.

The Tunnel Effect

The reason for the nonexistence of sharp, impenetrable boundaries for the confinement of electrons (and other small particles whose motion is governed by wave mechanics) can be made plausible as follows:

Imagine that we use a narrow plate condenser, consisting of two oppositely charged metal nets, as a barrier to the motion of electrons (Fig. 37.11). If the potential difference between the plates is high enough, the electrons will be turned back after penetrating a certain distance into the interspace between the plates. They will be turned back (reflected) and will proceed, after leaving the condenser, with the initial speed in the opposite direction.

This was the classical picture of the process. Now let us look upon it from the standpoint of wave mechanics. Our swarm of electrons of uniform momentum is accompanied by a de Broglie wave of definite wave length. According to wave mechanics the potential barrier behaves toward the de Broglie waves in a similar fashion as a glass plate behaves toward light; part of the wave is reflected and part of it is transmitted. The partial reflection of the wave implies that some particles are turned backward, and the partial transmission that some of the particles go through. *The striking difference between the wave-mechanical and classical theory consists in the prediction by wave mechanics of a transmitted wave even in cases where the kinetic energy of the incident particles is insufficient to carry them through the retarding field of the potential barrier.* This means that particles of a kinetic energy of 10,000 electron volts have a finite chance to pass through a retarding field between the condenser plates maintained at a potential difference of 100,000 volts. Such effects are appreciable only for fields whose intensity changes noticeably over the distance of one de Broglie wave length.

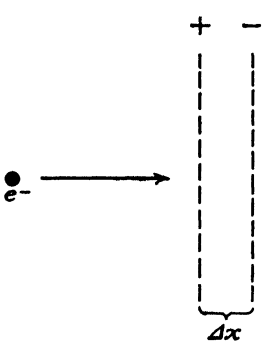


Fig. 37.11.

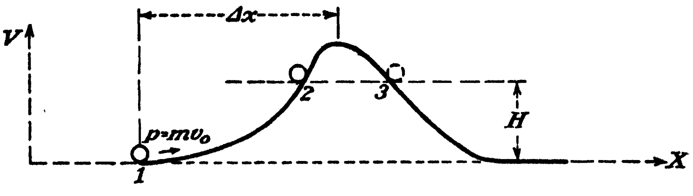


Fig. 37.12.

To clarify this statement, let us consider a mechanical analogy to the electrical case described above. Figure 37.12 shows a ball rolling without frictional energy losses up a steep hill which rises to its maximum height within a short distance Δx. We assume that the ball's momentum $p = mv_0$ would enable it to roll up to position 2 at the height H, according to classical physics, *but not to the top of the hill.* Wave mechanics predicts, however, that in a group of balls moving with the same momentum $p = mv_0$ a few may roll over the hilltop and be observed on the other side of its apex, as is shown by the dotted ball in position 3 of Fig. 37.12. In position 2 the kinetic energy of the ball is zero. Above position 2 it would become negative if the sum of kinetic and potential energy is to remain constant. Negative kinetic energy has no physical meaning, and hence its consideration will be tactfully avoided. The only way the ball in our example could get from position 2 to position 3 without violating the laws of classical physics would be through a tunnel dug for it through the hill. It is because of this mechanical analogy that the general phenomenon of penetration of potential barriers is referred to as the **tunnel effect.**

The wave mechanical theory shows that there is a finite probability for any particle, no matter how slow it is, to penetrate a potential barrier and that the probability of penetration increases rapidly as the kinetic energy of the particle approaches the value which would carry it to the top of the "potential hill," according to classical physics. In this latter case the probability of penetration is nearly 100 per cent. What can be said about a group of particles of uniform momentum applies as well to repeated attempts of the same particle to get through a potential barrier.

Thus we have seen that a potential barrier which, according to classical physics, would create a sharp boundary for the range of electrons of uniform momentum does not confine the electrons within a limited range but rather allows a certain percentage of the particles to leak through the barrier. In Chap. 38 we shall see how the theory of penetration of potential barriers is used to explain the radioactive decay of nuclei.

The Principle of Causality in Quantum Mechanics

The preceding discussion is apt to provoke alarming thoughts in the mind of the student. Even though quantum mechanics makes a very definite prediction about the percentage of the incident particles of a given momentum which will traverse a given potential barrier, no prediction is ventured about the fate of an individual particle. Some of them are assumed to pass, and others not, and there is no way of predicting which one will be the successful particle. But this makes no sense! Suppose we watch *identical particles* move with *identical velocities* toward the potential barrier of Fig. 37.11, one after the other. If particle 1 goes through, why should the identical particle 2 or 3 not do the same? Was not the entire structure of our science based on the assumption that the same thing which happens at one time under a given set of circumstances should happen at another time under the same set of conditions? Is it possible to give up this postulate and yet uphold the content and methods of physics?

Physics, as well as all the other natural sciences, is based on the validity of the **principle of causality** which Kant formulated as follows: **"Everything that happens presupposes something from which it follows according to a rule."** Heisenberg formulated it in a way which is more useful to us in the present consideration: **"If we know the present we can calculate the future."** Since a physical law is, in general, a rule of prediction, it is obvious that the existence of physical laws presupposes the validity of the principle of causality. The fact that wave mechanics does make correct statistical predictions about the behavior of elementary particles shows that the principle of causality holds in wave mechanics as long as one is concerned with the probability of events.

The failure of wave mechanics to predict the fate of an individual particle is explained by the argument that, according to the uncertainty relation, we cannot ascertain with unlimited precision the initial condition, *i.e.*, the instantaneous position and the momentum of the particle, and hence "since we do not know the present, we cannot calculate the future."

Thus quantum mechanics destroyed the dream of those who hoped to explain everything on the basis of Newtonian mechanics. This line of thought is best illustrated by a free rendition of a pronouncement by Laplace:

> If there were a being of superior intelligence [a so-called Laplacian Demon] who could ascertain with perfect precision the instantaneous positions and momenta of all particles in the universe, he could predict by mechanical computation with perfect accuracy all future events in the universe and he could equally determine all events of the past.

Since the initial condition of the system, as measured by the positions and momenta of the particles, cannot be determined with unlimited precision, quantum mechanics denies the possibility of predicting the course of events in the system. This led some physicists to rather radical conclusions about causality, such as the following statement of Born: "If it is impossible in principle to know all the conditions, that is, causes, of a process, it is empty talk to say that every event has a cause."

The renunciation of the principle of causality in considerations of the behavior of individual elementary particles is accepted by most modern physicists. There are, however, a few notable exceptions, such as Planck, Schrödinger, and Einstein, according to whom this is a transient state in the development of physics.

Quantum Mechanics and Biology

It is to be expected that, before long, quantum mechanical considerations will enter the field of biology as they have entered the field of chemistry. We know that structure and development of living matter depend on particles which approach molecular dimensions and which must interact according to the laws of quantum mechanics. Living organisms consist of minute organic units, the so-called *cells*, which in turn contain still smaller cell organs, such as cell nuclei, which in turn contain chromosomes—the site of genes, the carriers of heredity. The latter can be considered as giant molecules connected in chainlike fashion. The act of observing such minute structures involves exposing them to intense radiations (light, X rays, electrons, or other radiation). Experiments indicate, however, that radiation under certain conditions changes genes. Genes can be excited by absorbing discrete amounts of energy just as atoms were shown to be activated by discontinuous absorption of energy. Excitation of a gene is interpreted as consisting of a rearrangement of atoms in the gene; *i.e.*, an isomeric molecule is formed. This process is considered the explanation of gene mutations induced by X rays (see Chap. 34). But even at room temperature in the absence of X rays there is a certain probability of initiation of mutations due to exceptionally fast molecules in the thermally agitated environment of the gene, which may possess enough kinetic energy to excite the gene. Computations show, for instance, that at a temperature at which the average kinetic energy of the molecules is one-sixtieth of the excitation energy E of the gene, one would have to wait on the average 30,000 years in order to observe a gene excitation. This process of excitation can be considered as the mechanism of spontaneous gene mutation, which, as we know, is a very slow process. It was suggested that the evolution of the species is due to such spontaneous mutations induced by thermal molecular motion. The rate of spontaneous mutations is greatly speeded up by elevation of temperature.

APPENDIX

The Electron Microscope

This instrument is treated here, although its principle is not based on wave mechanics, because the appreciation of its high resolving power can be understood only on the basis of a wave mechanical consideration.

We have seen in Chap. 30 that the resolving power of a microscope is limited by diffraction phenomena so that object points which are separated by less than about 2,000 A cannot be seen as two separate points with the aid of visible light. Since the resolving power is increased by diminishing the wave length, it is possible to obtain somewhat higher resolving power by using ultraviolet instead of visible light. But the improvement is not spectacular. A very much higher resolving power could be obtained with X rays whose wave length is of the order of magnitude of 0.1 per cent of the wave length of visible light. But unfortunately, X rays cannot be focused as easily as light or as electrons.

The fact that the de Broglie wave length of moderately fast electrons, accelerated by about 60,000 volts, is of the order of magnitude of X-ray wave lengths suggests that the resolving power of an electron microscope would be hundreds of times as high as that of a light microscope if it were possible to focus electrons in a similar fashion to light. This idea proved feasible, and electron microscopes are now on the market whose resolving power approaches 20 A. It is possible to study with these instruments microscopic details a hundred times as small as those revealed by the optical microscope. We are now able actually to see large molecules, the genes in the chromosomes, and minute details in the structure of bacteria. This new visual aid promises to become as productive of revolutionary discoveries in biology as the original optical microscope.

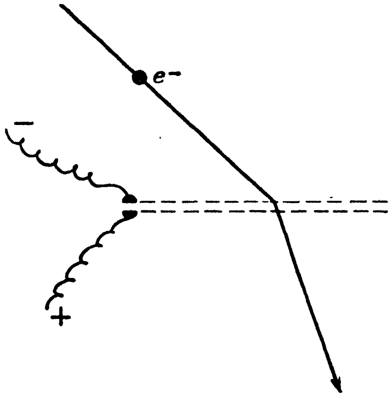


FIG. 37.13.

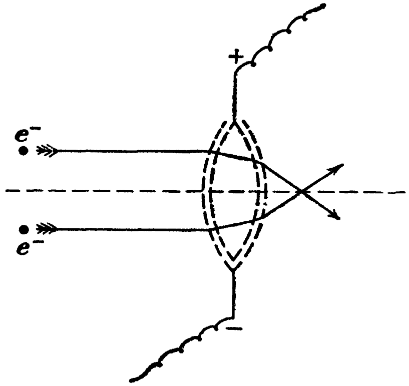


FIG. 37.14. An electrostatic electron lens.

Let us now see how it is possible to design electron lenses, *i.e.*, systems capable of focusing electrons, which are the basic elements of an electron microscope. In Chap. 28 we have seen how refraction of light is explained on the basis of the corpuscular theory. Figure 28.10 shows, as an illustration of the refraction of a light ray, the path of a corpuscle which is accelerated at the boundary as it enters an optically denser medium, *e.g.*, glass. A similar effect can be obtained with electrons if the "refracting surface" is represented by two closely placed condenser plates made of wire gauze between which a potential difference is maintained (Fig. 37.13). This principle can be used for the design of electron lenses consisting of an inner and an outer charged gauze such as shown in Fig. 37.14. Refraction of parallel electron beams by such a lens is quite analogous to refraction of light by a converging lens. The electrons are focused in fairly close agreement with the thin lens formula $(1/p) + (1/q) = 1/f$ (30.5). Practical electron microscopes use different types of electrostatic and magnetic electron lenses whose principle is, however, less simple to understand than the principle of the lens shown in Fig. 37.14.

The image formation by an electron microscope is analogous to the image formation by an optical microscope. The object (*e.g.*, bacteria or virus) is mounted on a very thin structureless membrane of cellulose and is introduced into the vacuum chamber (to avoid interference of gas particles with the motion of the electrons) of the electron microscope. An electron gun below the specimen "illuminates" it with electrons. After passing through the object, the electrons pass through a system of electron lenses which are capable of rendering a point image for a point source of electrons. The object, which modifies by absorption the intensity of the illuminating electron beam, can be considered as an electron source for every point of which an electron point image is formed on a fluorescent screen. Figure 37.15 shows a simplified self-explanatory scheme of the arrangement of the parts of an electron microscope.

The main disadvantage of the electron microscope in its present form lies in the necessity of introducing the object into a high vacuum, which quickly dries out living organisms, and in the subjection of the objects to an intense electron beam, which heats and usually kills microorganisms. Research on improvements is under way. Thus, for instance, it has been suggested to place the object (such as bacteria) between two thin membranes, which enclose it and prevent it from drying. Ardenne succeeded, using the conventional technique illustrated in Fig. 37.15, in following the cell division of a species of bacilli which are resistant to drying and can withstand the heating of the illuminating electron beam. He dried them before introducing them into the vacuum chamber and returned them to the nutritive solution after taking a microphotograph. By a repetition of this procedure he was able to follow the same bacterial cells through the process of cell division.

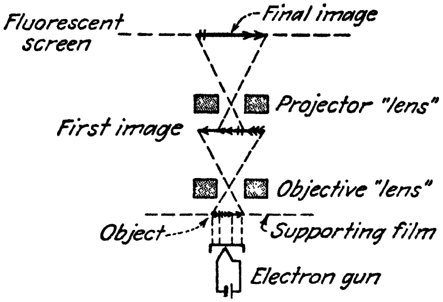


FIG. 37.15. Scheme of an electron microscope.

In most of the work which has been done, however, the organisms did not survive the process of observation. Nevertheless, a wealth of valuable information about the functions of microorganisms has been obtained from such investigations. Thus, for instance, structural details of some bacteria, such as flagellae, which had previously been suspected but not demonstrated, have been actually photographed, and many bacteria which are too small to be seen with the aid of an optical microscope have been made visible.

The magnification of the contemporary electron microscope is large enough to make individual virus particles visible. Thus one has found that the tobacco mosaic virus is rod-shaped and about 150 A wide and 3,000 A long. Influenza virus, on the other hand, is spherical and measures about 500 A in diameter.

In general it was found that particles of viruses which attack animals are larger than those which attack plants and that they show internal structure. For instance, the vaccina virus has a bricklike appearance, and one can distinguish in its interior five dense spots, which give it a dicelike appearance.

Interesting studies have been made on viruses which attack bacteria, the so-called *bacteriophages*. They look similar to spermatozoa, having a globular head with a straight tail. They attack bacteria head first or tail first and disintegrate them as was revealed by electron microscope photographs. Figure 37.16 shows coli bacteria attacked by bacteriophage anticoli.

Our ability to see virus particles is of the utmost biological interest, since they are considered to be large molecules of living substance which constitute a transitional state from inanimate to animate matter. It is hoped that the study of these particles will bring us closer to the answer of the question: "What is life?"

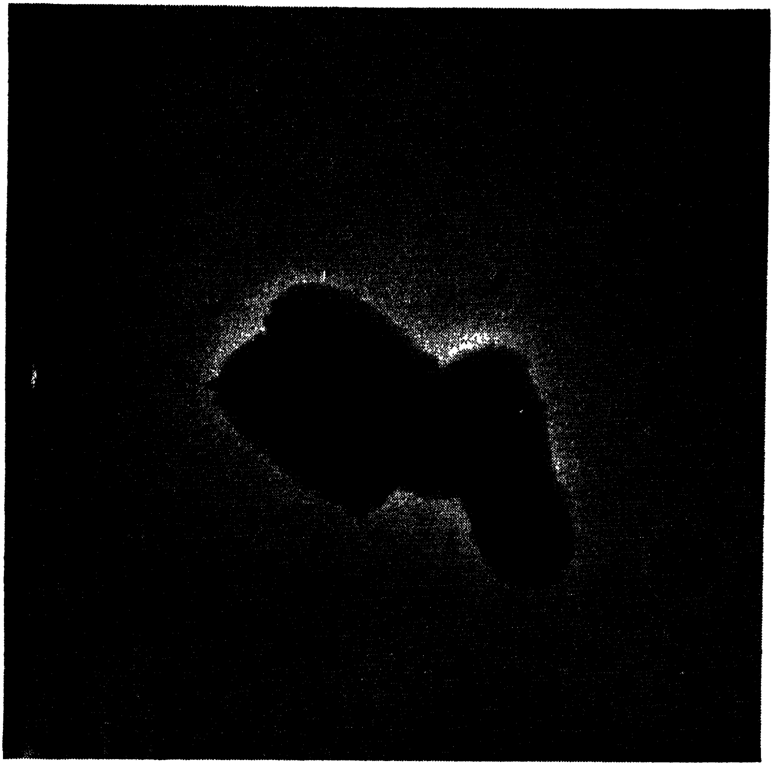


FIG. 37.16. Electron microscopic photograph showing bacteriophage anticoli attacking coli bacilli (20,000-fold magnification). (*After Luria and Anderson. Courtesy of V. K. Zworykin.*)

As stated before, the limit of resolution of a modern electron microscope is 20 A. Taking into consideration the fact that there are organic molecules, such as the biologically important proteins, whose size is considerably greater than 20 A, we see that this resolving power enables one to discern the shape and thus to study the structure of large organic molecules. Ardenne succeeded in photographing molecules of hemocyanin, which are about 200 A long. Thus the electron microscope enables us actually to see large molecules and to study their structure in a way never dreamed of by chemists in the past.

QUESTIONS AND PROBLEMS

1. If you were to observe scattering of X rays in various directions, would you observe a greater change in frequency for those X-ray photons which are scattered at right angles to the direction of the original beam than for those which are scattered back toward the source?

***2.** Can you conceive of a process in which the wave length of an electromagnetic radiation is shortened by the scattering process?

3. *a.* Compute the mass of a "violet photon" of $\lambda = 4{,}000$ A.

b. Compute its momentum.

c. How fast would an electron have to move in order to have the same momentum?

4. What would be the uncertainty in the position of an electron if we were to specify its momentum with unlimited precision?

5. Compute the de Broglie wave length of the following:

a. An electron moving with speed $v = 10^9$ cm/sec.

b. A proton moving with speed $v = 10^9$ cm/sec.

c. A mass of 1 gm moving with speed $v = 10^9$ cm/sec.

d. A mass of 1 gm moving with speed $v = 1$ cm/sec.

6. What postulates of classical physics do we renounce when we accept wave mechanics?

7. What postulates do we adopt in wave mechanics?

8. In what way are we justified in speaking of a distribution of charge density of the electron cloud in Schrödinger's atomic model?

9. Compute the uncertainty in the momentum of a proton whose position is localized within $\Delta x = 0.01$ mm. Estimate also the uncertainty for its velocity.

10. Determine the location of diffraction minima for a beam of U^{235} atoms passing through a slit of width $\Delta x = 0.001$ mm with a speed of 100 m/sec. Assume a screen 10 m away.

11. Explain the principle of the electron microscope.

12. What would be the advantage, if any, of a proton microscope as compared to an electron microscope? (A proton is a hydrogen nucleus.)

BIBLIOGRAPHY

Schrödinger, E., "What Is Life?" Macmillan, 1945.
Bohr, N., Light and Life, *Nature*, Vol. 131, pp. 421, 457, 1933.
Burton and Kohl, "The Electron Microscope," Reinhold, 1942.

CHAPTER 38

SPONTANEOUS TRANSMUTATION OF ELEMENTS

a. By which two numbers is a nuclear species specified?

b. What was the content and basis of Prout's hypothesis?

c. How was Prout's hypothesis modified by modern findings?

d. What data enable us to make statements about the relative stability of nuclei?

e. How do we measure the intensity of radioactive radiations?

f. How does the phenomenon of radioactivity affect the traditional definition of a chemical element?

g. What is the source of the energy evolved during the process of decay of radium?

h. How can you distinguish and identify the radiations emanating from a mixture of radioactive elements?

i. What is the principle of the Geiger counter

j. How does the Wilson chamber work?

k. Explain in detail what (1) observations and (2) reasoning led to the neutrino hypothesis.

l. Why do not all nuclei of radium disintegrate at the same time?

m. What is the "decay constant"? What does it measure?

n. What is the "half life" of a nucleus? What is the relation between the half life and decay constant?

o. Is there a relation between the energy of emitted alpha particles and the rate of decay?

p. Describe Gamow's theory of alpha decay.

q. How does quantum mechanics explain the statistical nature of radioactive decay?

r. What is a meson?

s. What happens to a meson eventually in a vacuum?

Relations between Nuclear Species

It was noticed early in the nineteenth century as a curious fact that the atomic weights of all elements were nearly integral multiples of the atomic weight of hydrogen. Prout attempted to explain this simple relationship between the atomic weights by assuming the atoms of all elements to be composed of hydrogen atoms. This hypothesis, if acceptable, would be a great step toward simplification of our views on the constitution of matter, as it would reduce the complexity of the material world by assuming one fundamental substance instead of some 100. This hypothesis was, however, short-lived. It was soon discovered that not all atomic weights were nearly integral multiplies of the atomic weight of hydrogen; thus, for instance, the atomic weight of chlorine was found to be 35.457, and many other exceptions were found. This led chemists to reject Prout's idea. Investigation of chemical atomic weights by means of the mass spectroscope (in 1913), however, revealed that elements consist of isotopes† whose masses are very nearly integral multiples of the mass of the hydrogen atom. The nonintegral values of the atomic weight of elements such as chlorine were explained as being average values in a mixture of isotopes of nearly integral atomic weights (see Chap. 23).

This led to a revival of Prout's hypothesis, but not without essential modifications. Since the mass of an atom is almost entirely concentrated in its nucleus, Prout's assertion, if put in the terminology of the nuclear atomic model, could be interpreted as stating that the nucleus of any element is composed of hydrogen nuclei. If this were the case, the nuclear charge (and hence the atomic number) of each element would have to be equal to its atomic weight. A glance at the periodic table shows, however, that this is not the case. For the lighter elements the atomic weight is about twice as great as the atomic number, and for heavier elements it is more than twice as great. Prout's hypothesis, moreover, fails to explain the existence of isotopes, *i.e.*, of elements of identical nuclear charge but different nuclear mass.

But the situation does not seem to be quite hopeless. Could we not assume an appropriate number of electrons in the interior of the nucleus so as to neutralize the positive charge of some of the hydrogen nuclei? This would enable us to obtain the desired values of nuclear charge for a given nuclear mass. Unfortunately, however, according to wave mechanics it is quite unreasonable to assume the existence of an electron inside a nucleus for reasons which will be mentioned later in this chapter. We shall therefore adopt provisionally a more radical hypothesis, either assuming the existence of negative particles of electronic charge which are different from electrons (**mesons**)‡ and whose existence inside a nucleus is not inconsistent with wave mechanics or postulating the existence of neutral particles (**neutrons**) of a mass approximating the mass of the hydrogen nucleus. These "neutrons," together with hydrogen nuclei (**protons**), could be then considered as the building blocks (**nucleons**) of all nuclei. We leave it an open question, for the time being, whether these neutrons are elementary indivisible particles or a very close combination of a proton with an electron or a "meson."

The following terminology and symbols will be used in the subsequent description of atomic nuclei:

Atomic number Z = nuclear charge = number of protons in the nucleus.

---

† Isotopes were discovered previously in studies of radioactive decay.

‡ Such particles have actually been observed. The mass of the *mesons* is larger than the electron mass. More will be said about them later on.

Atomic mass M:[†] The mass of an atom in terms of *atomic mass units* (amu). The atomic mass unit is defined as one-sixteenth of the mass of an oxygen O^{16} atom; it is very nearly the mass of a proton.

Mass number A: The nearest integer to M. A has the theoretical meaning of the total number of nucleons in the nucleus.

We adopt the hypothesis that the nucleus of an element of mass number A and atomic number Z consists of A nucleons of which Z nucleons are protons and $A - Z$ nucleons are neutrons. *According to this, isotopes are nuclear species having the same number of protons but different numbers of neutrons in the nucleus.*

The mass number and atomic number of an element are symbolized as follows: In the symbol ${}_ZS^A$, S is the chemical symbol of the element, Z the atomic number, and A the mass number. For instance,

$$ {}_3\mathrm{Li}^7;\ {}_5\mathrm{B}^{10};\ {}_7\mathrm{N}^{14};\ {}_{24}\mathrm{Cr}^{50};\ {}_{92}\mathrm{U}^{235};\ \ldots $$

Our hypothesis on the constitution of atomic nuclei encounters immediately a serious difficulty: It seems that the mass of any nucleus should be *exactly* equal to Z times the mass of the proton plus $(A - Z)$ times the mass of the neutron. We can therefore undertake to compute the neutron mass from, say, the mass difference of the two lithium isotopes ${}_3\mathrm{Li}^6$ and ${}_3\mathrm{Li}^7$. We get for the difference of their masses:[‡] $7.01816 - 6.01692 = 1.00124$ amu (atomic mass units), which we take to be the mass of the hypothetical neutron.[*] Now we may try to predict, for instance, the mass of the deuterium nucleus (heavy hydrogen ${}_1\mathrm{H}^2$), which should consist of a neutron and a proton, according to our hypothesis. We get from our data $m_p + m_n = 1.00757 + 1.00124 = 2.00881$ amu. The experimentally determined mass of the deuterium nucleus is, however, $m_D = 2.01417$.

Stability of Nuclei

Does this disagreement between prediction and experiment mean that our hypothesis about nuclear structure is untenable? Not necessarily. It may indicate that our method of estimating the neutron mass was wrong or that our method of computing the nuclear mass from the neutron and proton masses was incorrect. Actually it can be shown that both steps in our considerations were wrong for the following reason: Imagine two elementary particles which attract each other. It makes no difference in this consideration whether the forces of attraction are electrical, magnetic, gravitational, or of some other kind. The system consisting of these two particles (which are imagined to be at rest relative to each other in subsequent considerations) will possess potential energy, which will be greater the farther apart the particles are. A system A of two such particles placed very closely together will possess less potential energy than a system B in which these two particles are far apart. If system A has been formed from system B, energy must have escaped somehow in this process (for instance, in form of electromagnetic radiation). Now according to Einstein's law ($E = mc^2$), the loss of energy by our system should be accompanied by a loss of mass. If the energy loss is given by ΔE, the loss of mass should be $\Delta m = \Delta E/c^2$. In other words, the combined mass of the two particles should be less when they are close together than when they are infinitely far apart; *i.e.*, the mass of the "double particle" must be smaller than the sum of the masses of the individual particles.

[†] The mass m of the corresponding nucleus is obtained by subtracting from the atomic mass M the mass of the planetary electrons. The mass of an electron in atomic mass units is $m_0 = 0.00056$.

[‡] In this calculation it is immaterial whether we use the masses of the atoms or the masses of the nuclei, since both atoms contain the same number of electrons, whose masses cancel out in the subtraction. The values used above are the atomic masses.

[*] The mass thus found for the neutron is wrong! The mass found for the neutron experimentally is $M_n = 1.00890$ amu (p. 806).

Hence, from Einstein's law, we infer that the mass of a nucleus whose nucleons are bound by mutual attraction should be smaller than the sum of the masses of the individual nucleons. We shall call the difference between the nuclear mass and the sum of the masses of the separate nucleons the **mass deficit** (Δm).†

We see from this that our attempt to obtain the mass of the free neutron from the mass difference of two nuclei differing by one neutron could not succeed. We also see that, even if we had the correct masses of the neutron and the proton, we could not obtain from them the mass of the deuterium nucleus or any other nucleus by simple addition. The mass of a stable nucleus should be smaller than the sum of the separate nucleon masses, but we cannot say how much smaller without knowing the "binding energy" (potential energy loss) of the nucleons in the nucleus.

As will be shown later, neutrons have been actually discovered and the neutron mass has been determined accurately. This makes it possible to compute the mass deficit of the various nuclei. Such computations are of interest, since *the mass deficit can be considered to be a measure of the stability of the nucleus.* The product of the mass deficit (in grams) and the square of the velocity of light (Δmc^2) can be considered to be the **binding energy** of the nucleus, *i.e.*, *the amount of energy which we have to supply to the nucleus in order to separate the nucleons from each other.* The greater Δm the larger is the binding energy and the more stable is the nucleus.

---

† This is not the same thing as the "mass defect" used in some other books which is defined as the difference ($M - A$) between the atomic mass and mass number.

If we divide Δm by the mass number A, we obtain a quotient which can be considered as the *binding energy per nucleon*. When we plot $|\Delta m/A|$ as a function of the mass number A, we obtain a curve which rises rapidly to a maximum in the vicinity of $A \simeq 60$ and drops gradually from there on. This means that the binding energy per nucleon and, hence, *the stability of the nucleus are highest for nuclei of medium atomic weight* and that the stability gradually decreases with increasing atomic weight.

Discovery of Unstable Nuclei

The fact that the heaviest nuclei turn out to be among the least stable ones leads us to expect that we might be able to observe their decomposition into individual nucleons or possibly into larger fragments. The discovery of spontaneous disintegration of heavy nuclei was actually made at the beginning of this century, before the concept of "nucleus" had been formed, in an investigation which was not at all concerned with structure of matter. The newly discovered phenomenon was called **radioactivity.** This discovery was made a few months after the discovery of X rays as a direct consequence of the latter. Roentgen reported X rays to emanate from a cathode-ray tube issuing from the glass portion which was hit by the cathode rays. This portion was made fluorescent by the cathode rays, and the question was asked if there was some causal relation between the fluorescence of the glass and the emission of X rays by it. Experiments were made to determine whether X rays were emitted from substances which have been made to fluoresce or to phosphoresce by irradiation with ultraviolet light. Several authors reported that some fluorescent substances gave off penetrating rays similar to X rays after exposure to sunlight. H. Becquerel decided to investigate these reports and prepared to repeat the experiments with the fluorescent double sulfate of potassium and uranium. His plan was to expose the salt to the radiation of the sun and then to place it on a photographic plate after wrapping the salt in black paper in order to see whether a penetrating, photographically active phosphorescent radiation was emitted by the salt capable of passing through the black paper. Having to wait several days for sunshine, Becquerel performed a control experiment in which the above procedure was followed, except that the salt was not previously exposed to sunlight. Much to his surprise he found the effect which he expected to find with preirradiated salt only. A penetrating radiation capable of passing not only through black paper but also through a dense wrapping of tin foil issued from the salt and blackened the photographic plate. This occurred even with samples of the salt which were especially synthesized in the dark and never exposed to daylight.

It was soon found that this "radioactivity" was not a specific property of the above-mentioned fluorescent salt but that it was a general property of all uranium compounds, even of those which do not show fluorescence, and that the pure uranium element is also radioactive. This finding was soon followed by the discovery that thorium and its compounds behaved similarly to uranium.

It was established that the intensity of the penetrating radiation issuing from a mineral sample was proportional to its content of uranium or thorium. The measurement of the intensity of radioactivity used in the above investigation was based on the observation that the penetrating radiations of uranium and thorium had the property of ionizing air. Figure 38.1 shows an electroscope with an attached ionization chamber consisting of two parallel metal plates P_1 and P_2, which can be charged by closing switch S. (B is a voltage supply.) The divergence of the leaves measures the potential difference between the plates. When a radioactive mineral M is introduced between the plates, the air between them is rendered conductive by ionization and the divergence between the leaves begins to diminish as soon as the switch S is opened. The rate of convergence of the leaves is taken as the measure of the intensity of ionization, which serves as the measure of radioactivity. The rate of convergence can be measured by determining the time it takes the deflection of the leaves to be reduced by 50 per cent.

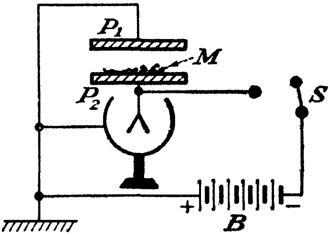


FIG. 38.1. Measurement of radioactivity of a mineral sample.

Measurements of this type were performed on a great number of minerals, and attempts were made to account for the observed intensity of radiation by the uranium and thorium content of the mineral. In the course of such studies, the Curies observed that the radioactivity of the mineral "uranium pitchblende" was several times as intense as could be accounted for by its uranium content. They concluded that the observed excess radioactivity was due to traces of an unknown, highly radioactive ingredient of that mineral. After an exceedingly tedious procedure involving the processing of tons of pitchblende, they succeeded in isolating a few milligrams of the highly radioactive component of the mineral, a new element **radium**, whose intensity of radioactivity turned out to be more than a million times that of uranium. Its atomic weight was found to be 226, and its chemical properties indicated that it belongs to the group of alkaline-earth metals such as calcium, strontium, and barium.

Radium can be shown to be an element by heating it to very high temperatures at which all known compounds decompose, and showing that it preserves its chemical identity. And yet if we enclose a piece of freshly prepared, purified radium in an evacuated and subsequently hermetically sealed vessel, we soon find that a gaseous atmosphere begins to accumulate in the originally evacuated container. Spectroscopic analysis reveals that *helium* is one of the gases which seem to emanate from the enclosed piece of radium. The additional spectral lines are ascribed to another gas (radium emanation or **radon**) which can be separated from helium by cooling and adsorption on charcoal. Radon turns out to be a member of the same family of elements as helium, namely, of the inert gases.

Spontaneous Transmutation of Elements

The evolution of two new elements from radium is a discovery of the greatest importance. It demonstrates that the original definition of a chemical element as an immutable substance is inadequate.

Radium of mass number $A = 226$ changes into radon of mass number $A' = 222$ and helium of mass number $A'' = 4$, so that $A = A' + A''$. Previous to the discovery of radium Becquerel had demonstrated a similar transformation of uranium into other elements (see Fig. 38.11 showing chains of transformation of uranium and its "daughter elements"). The examples of the transmutation of radium and uranium were eventually expanded by many other instances of spontaneous transformation of heavy elements which thus manifest their instability.

Calorimetric measurements indicate that the transmutation of elements is accompanied by a release of energy which is enormous in comparison with energies liberated in chemical reactions. Each gram of radium develops about 140 cal/hr owing to decomposition of some of its atoms, and it can be computed that the total amount of heat which would be released by the 1 gm mass in a process of complete decomposition into radon and helium would be 3.7×10^9 cal. *This figure is about one million times as large as the heat liberated in the explosive formation of* 1 *gm of* H_2O *from hydrogen and oxygen.* Where does this enormous energy come from? In the early days of the discovery of radioactivity the spontaneous generation of energy by radioactive elements, whose transmutation was not as yet known, threatened to shake confidence in the general validity of the law of conservation of energy. Later Elster and Geitel suggested "that the atom of a radioactive substance giving off energy passes, in the manner of a molecule, from an unstable combination into a stable condition." This view was adopted and reinforced by Rutherford and Soddy, who demonstrated the transformation of radium into radon and helium. The source of this energy release was finally clearly recognized after Einstein established the law of transformation of mass into energy. *It was shown in recent years through very many exact experiments that, whenever one nucleus decays into two "daughter nuclei," the sum of their masses is smaller than the mass of the "parent nucleus." The loss of mass accounts precisely for the appearing energy according to Einstein's mass-energy law* $E = \Delta mc^2$. *This energy is observed partly as kinetic energy of the daughter nuclei which fly apart in the violent nuclear fragmentation.* In the above reaction $\text{Ra} \rightarrow \text{Rn} + \text{He}$, we assume the mass of the radium nucleus to be slightly larger than the combined masses of the nuclei of radon and helium to account for the heat evolved by decaying radium.

Numerical examples illustrating the transformation of mass into energy in nuclear transmutations will be presented in Chap. 39.

Radiations

A piece of radium is heated primarily by absorbing the ionizing radiations which its disintegrating atoms emit. It has been shown that

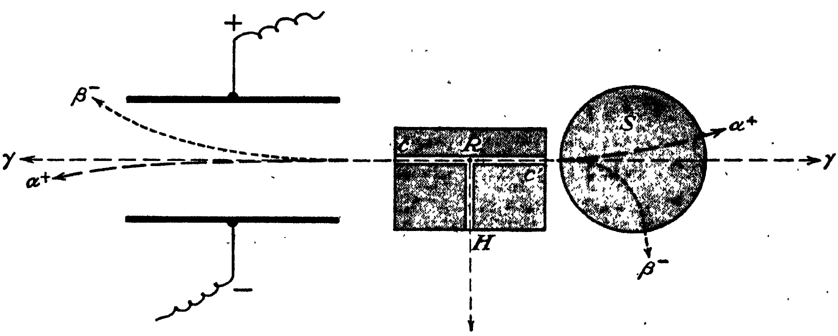


FIG. 38.2. Analysis of radiations from a radioactive sample.

the energy liberated in the decay of radium and of other radioactive elements is transferred through space in form of different radiations, namely, as kinetic energy of nuclei or electrons and as energy of emitted photons. An analysis of these radiations can be accomplished as follows: Figure 38.2 shows a lead cylinder through which a thin hole cc' is drilled and another hole H at right angles to it. A piece of radioactive material R consisting of a mixture of uranium, radium, and many other radioactive elements and their decay products is placed at the intersection of the channels H and cc'. The radiation which escapes through H

renders the air conductive and fluorescent and heats solid objects upon which it impinges. It can be shown that this radiation is a mixture of three types of radiation—*α rays, β rays,* and *γ rays*—by allowing the radiation escaping from c' to pass through a magnetic field [which is perpendicular to the plane of the paper, only one of the two poles (S) being shown] and (or) allowing the radiation escaping from c to pass through an electric field in a vacuum. In both cases the original beam is split into three beams, which are labeled α, β, and γ in the figure.

1. *γ rays:* An undeviated beam which may consist of neutral particles or of photons or both.

2. *β rays:* Negatively charged particles which are deviated easily.

3. *α rays:* A beam whose deviation indicates that it consists of positive charges. The smallness of the deviation suggests that they have a large momentum.

1. **The γ rays** are distinguished by their penetrating power. They can penetrate several centimeters of lead without being completely absorbed. They ionize the air and differ from X rays only by being more penetrating. Sending a thin beam of γ rays through a crystal produces a Laue diffraction pattern similar to the one obtained with X rays. This allows the determination of their wave length, which turns out to be of the order of magnitude of 0.01 A, or about one-tenth as large as that of commonly used X rays. *γ rays are X rays of exceedingly short wave length. They can be generated artificially in betatrons and X-ray tubes at voltages of several million volts.*

2. **The β rays** can be identified as electrons by determination of their specific charge e/m. Here an initial surprise was encountered, however. e/m turned out to be somewhat smaller than that of slow cathode-ray electrons.

This finding was clarified by experiments of Kaufmann and by the relativity theory of Einstein. Kaufmann found that e/m of electrons decreases as their velocity is increased. At speeds very close to the velocity of light e/m may be a small fraction of its value at slow speeds. Assuming the charge to be independent of speed,† *we interpret this result as indicating a variation of mass with speed.* This is consistent with Einstein's mass-energy law. At higher speeds electrons possess a greater kinetic energy and hence, according to Einstein's law, a greater mass. *Thus we conclude that the β rays consist of very fast electrons (moving in some instances with speeds exceeding* 99 *per cent of the speed of light) whose mass is increased as a result of their high speed.*

---

† This is not the only admissible assumption. Vannevar Bush has worked out a theory in which the charge is assumed to vary with speed.

The penetrating power of the β rays is quite small as compared with that of γ rays. Their intensity is about halved by passing through 1 mm of aluminum.

3. **The α rays** are the least penetrating of the three types of radiation. Their intensity is about halved by passing through 10^{-2} mm of aluminum.

Rutherford and Soddy determined the nature of the α **particles** (the corpuscles of which the α rays consist) as follows:

They sealed a piece of radium in a very thin glass tube g, whose wall was so thin that α rays could penetrate it (Fig. 38.3). This tube was placed inside a gas-discharge tube T containing a rarefied gas, say hydrogen. Initially the hydrogen spectrum was obtained when passing an electric discharge through the tube. But after several days, lines of helium also appeared in the discharge spectrum. The two investigators interpreted this experiment by assuming that the α particles, which are emitted by the piece of radium R and pass through the wall of the thin glass tube g into the discharge tube T, are nuclei of helium. If this assumption is correct, we should be able to show that the charge on an α particle is equal to two elementary charges, since we have to strip two electrons from a helium atom in order to obtain a helium nucleus. The charge q on an individual α particle can be determined by measuring q/m by electrostatic and magnetic deflection, since m is known for a helium nucleus. The result confirms the assumption that *α particles are doubly positive helium ions, i.e., nuclei of helium.*

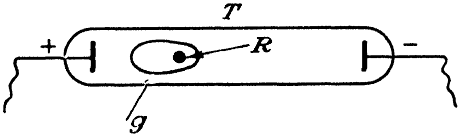


FIG. 38.3. Experiment to determine the nature of α particles.

Detection of Individual Elementary Particles

The rapid advance of our knowledge about atomic nuclei was to a large extent due to the invention of methods of observation which made it possible to detect individual elementary particles, such as electrons and α particles. We shall discuss the following four methods of this kind: the scintillation counter, the Geiger-Müller counter, the Wilson cloud chamber, and the photographic emulsions for nuclear studies.

1. **The Scintillation Counter.** In his famous α-scattering experiments Rutherford used a zinc sulfide screen to detect individual α particles which cause the screen to emit a faint flash of light when they hit it. This is a very strenuous method of observation. Recently very sensitive photoelectric cells have been developed, the so-called "photomultiplier"

tubes, which make it possible to obtain a recordable electric impulse for the weak light flashes produced by individual α particles. Instead of zinc sulfide various organic fluorescent materials are used, such as anthracene.

2. The Geiger-Müller Counter. If somebody had told a physicist in 1900 that it would be possible in 1920 to detect an individual electron or an ion, he probably would have been incredulous. After all, detection implies some kind of a reading or signal, such as the deviation of an instrument pointer or, say, a click of a loud-speaker membrane. How could a tiny elementary particle provide enough energy to move an instrument pointer or a loud-speaker membrane? The answer is that it cannot do it directly, but neither can we throw a rifle bullet with a speed of 500 m/sec. We can, however, press a trigger which releases a chemical chain reaction that provides the required energy. Similarly, one has succeeded in making an individual electron or ion "trigger" a powerful electric gas discharge. Figure 38.4 shows a *Geiger-Müller counter*, which allows the detection of individual ions, electrons, or (indirectly) photons. The axial wire A, which is sealed in a cylindrical glass tube T, is maintained at a high positive potential (about 1,000 volts) with respect to the cylindrical negative electrode C. A small opening in the cylinder in front of the window W allows the radiation to enter the cylinder space. The window W may be just a section of the glass tube T where the glass has been blown out to a small thickness, or it may be a thin foil of aluminum or some other material which absorbs the radiation as little as possible. The gas in the tube is usually maintained at atmospheric pressure, but if necessary, pressures down to a few centimeters of mercury can be used.

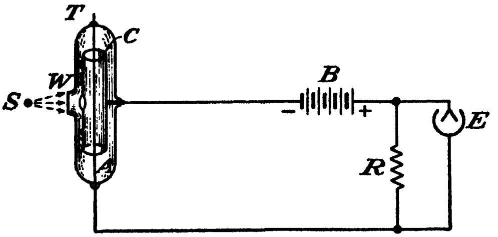


Fig. 38.4. Circuit of a Geiger-Müller counter.

Let us now suppose that an electron from the source S enters the counter. It will be accelerated toward the wire in the field between the cylinder and the wire, and as it moves, it will ionize gas atoms by collision. After the first collision there will be three particles: two electrons, which are rapidly accelerated and will cause further ionization, and a heavy gas ion, which drifts comparatively slowly to the cathode. After these two electrons have each ionized another atom, there will be four electrons (plus two additional ions). It is easy to see that after several more "generations" the number of electrons and ions generated by the original electron will increase enormously. This process reminds one of the formation of an avalanche.

Thus in a tiny fraction of a second the electron which entered our counter cylinder creates an "electric avalanche," *i.e.*, a relatively intense current, which flows through the resistance R, creating a voltage drop across it which is recorded as a momentary "kick" on the deflecting electrometer E.† This voltage drop can also be amplified (by electronic tubes) and made to actuate a mechanical counting device recording the number of such impulses and thus counting the number of electrons entering the counter.

The resistance R has also another important function. It stops the gas discharge initiated by an electron and prepares the counter to record the next electron. The voltage drop across R diminishes by iR the battery voltage V applied to the counter. At a certain value of the instantaneous current i, the potential difference across the counter $(V - iR)$ becomes insufficient to maintain the gas discharge and the counter is "quenched."

A similar process takes place when an ion enters the counter (instead of an electron). Since ions are more readily absorbed than electrons, the window must be particularly thin (very thin mica windows can be used) when the counter is used to count, for instance, α particles.

Such a counter can also be used with photons. In the case of γ rays and X-ray photons no window is necessary. The penetrating quanta pass through the glass wall and through the metal cylinder C, ejecting from the latter photoelectrons or recoil electrons due to Compton effect. These electrons are then counted. Not every incident quantum is absorbed in the cylinder, and not every absorbed quantum gives rise to an electron which happens to enter the cylinder space. Thus the efficiency of the counter in detecting photons is not very high.

Photons of visible light can also be recorded if a metal of low work function (such as alkali metals) is deposited on the inner cylinder wall. This type of "photoelectric quantum counter" was developed by Rajewski in order to check the claims of the biologist Gurewitch, who believed that he had discovered "mitogenic rays," *i.e.*, supposedly an ultraviolet electromagnetic radiation, emanating from growing onion roots and other growing tissues. All attempts to confirm physically the existence of such a radiation have failed so far.

† Electronic counting circuits are used in practice.

The Geiger counter records infrequent ionizing events in the absence of a source of ionizing particles. These "background counts" are partly due to traces of radioactive elements in the environment. If all the counts were due to it, we should expect the "background" to disappear were we to ascend with a counter in a balloon. Much to his surprise Hess found that the intensity of ionizing radiations increases rapidly with altitude. These observations led to the discovery of the so-called **cosmic rays,** a highly penetrating radiation consisting of a variety of particles (see pages 808, 811), some of which originate in our atmosphere while others ("primaries") appear to come from interstellar space. The origin of primary cosmic rays is as yet unknown.

3. **The Wilson Cloud Chamber.** Another device which has led to the discovery of several new elementary particles and new processes is the *Wilson cloud chamber.* It is perhaps no mere matter of chance that this ingenious method was invented by a Londoner. London is well known for its fogs, the formation of which is favored by great quantities of tiny coal particles, which are floating in the London sky as a result of incomplete combustion. These coal particles do not constitute fog, but they rather "invite" its formation. Fog consists of tiny water droplets which, owing to their smallness, fall with an imperceptibly small terminal velocity. Condensation of fog droplets occurs when there is a sudden drop in the temperature of the atmosphere. Such a temperature drop usually occurs when the barometric pressure drops. It is a result of the adiabatic expansion of the air, which cools as it does work in expanding.

It has been found that air which is supersaturated with moisture may be "undercooled" considerably without the formation of fog if it contains no dust particles. The slightest amount of dust will, however, precipitate the formation of fog, since water droplets have a tendency to condense on so-called *cores of condensation. Dust particles, or ions, can serve as cores of condensation.*

Wilson conceived of the following idea: Imagine an atmosphere of purified air saturated with water vapor. When an α particle passes through this atmosphere, it will ionize many atoms along its path, so that its trajectory is littered with "debris" consisting of positive and negative ions.† *The track of the α particle will be marked by a linear "cloud" of ions.* If we now suddenly allow the atmosphere to expand adiabatically, it will cool and water droplets will condense on the ions, forming a linear cloud of droplets. This cloud could be made visible by lateral illumination. This idea proved eminently successful.

---

† Many of the liberated electrons may be captured by atoms and form negative ions.

Figure 38.5 shows the design of a simple cloud chamber. It consists of a transparent cylindrical portion *Cy* covered by a glass plate *g* and a conical portion *Co* to which a rubber bulb *RB* is attached. The rubber bulb and part of the chamber are filled with water *W*, the surface of which serves as a piston of the expansion chamber. When the bulb *RB* is squeezed, the water rises in the chamber *Cy*, compressing and heating the air above. We wait about a minute until the air cools (owing to heat exchange with its environment) and becomes saturated with water vapor. Then we suddenly release the bulb *RB*. The compressed air in the chamber expands adiabatically and cools. If a source *S* of α particles is mounted inside the chamber, we see cloud tracks diverging from *S*.† The illumination is accomplished by a lateral source of light *L*. The eye of the observer is located at *O*. An actual photograph of α particle tracks is shown in Fig. 38.6. It represents the tracks of α particles emitted by thorium *C* and thorium *C′*. We see that two groups of α particles are present each characterized by a definite range, *i.e.*, distance traversed in air.

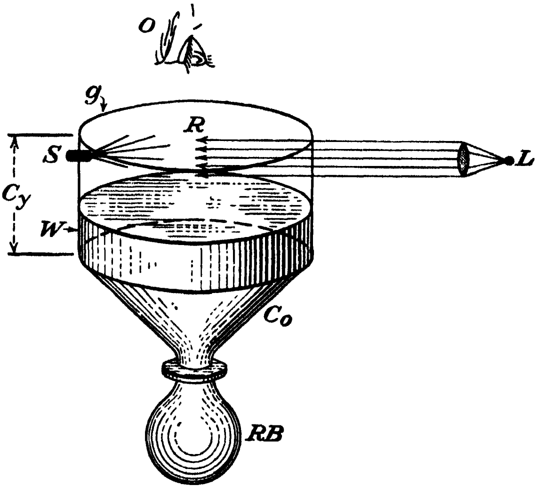


FIG. 38.5. A simple Wilson cloud chamber. (The cloud tracks are made visible as bright lines on a dark background.)

Tracks of β ray electrons can also be made visible by this method, but they are much weaker. Figure 38.7 shows the path of a beta ray electron next to an alpha particle track. We see that, whereas the latter is short, straight, and dense, the former is bent, long, and faint, enabling one to distinguish individual droplets.

---

† There is a provision for sweeping ions out of the air space by means of an electric field before each expansion. This arrangement is not shown in the figure.

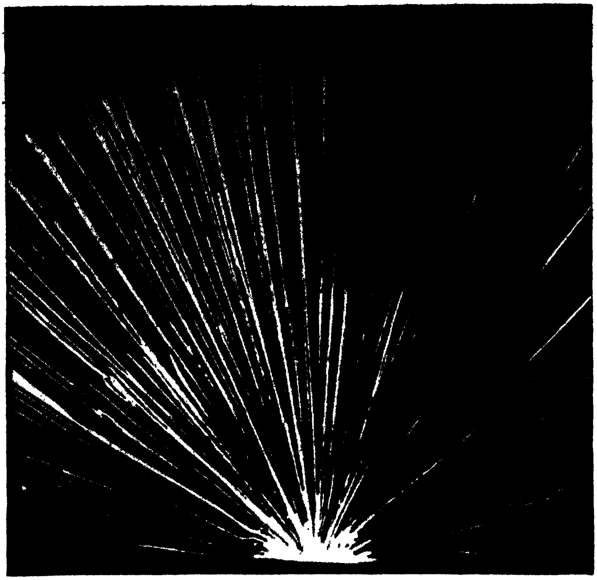


Fig. 38.6. Cloud-chamber tracks of α particles emitted by a mixture of thorium C and thorium C'. (*Courtesy of C. T. R. Wilson.*)

Gamma rays and X rays do not give rise to distinct tracks of their own path. Their path is marked by irregularly bent tracks of electrons which are knocked out of the gas atoms by Compton effect and by the photoelectric effect. Figure 38.8 shows the tracks of the secondary electrons produced by a horizontal beam of gamma rays.

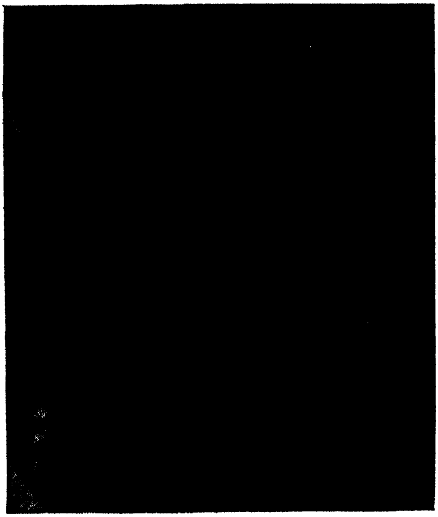


Fig. 38.7. Cloud-chamber tracks of an α particle (*left*) and an electron (*right*). (*Courtesy of W. E. Hazen.*)

4. **Photographic Emulsions.** Lately photographic emulsions have been developed which are sensitive to fast ions such as, for instance, α particles and even to fast electrons. The track is marked by a sequence of black silver grains which are formed from silver bromide grains by passing ionizing particles.

Figure 38.9 shows the following events taking place *inside the photographic emulsion.* In the upper left-

FIG. 38.8. Cloud-chamber photograph showing ionization of air by γ rays. (*Courtesy of C. T. R. Wilson.*)

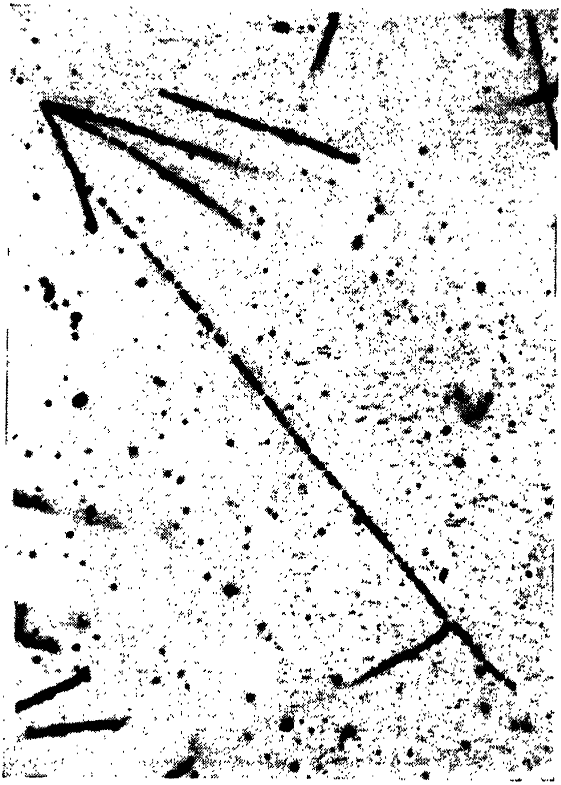


FIG. 38.9. Collision of an α particle with a proton in a nuclear photographic emulsion. (*From Powell and Occhialini, "Nuclear Physics in Photographs," Oxford, Clarendon Press.*)

hand corner the three prongs are the tracks of α particles which emanated at different times from a radio-thorium nucleus which underwent three successive α disintegrations. The long track originating from the α track on the left is the track of a proton which has been hit by the α particle. We see how the density of the proton track increases toward its end. At the beginning of the proton track the silver grains are less densely spaced than in the α particle track (why?).

The photographic emulsions are particularly suitable for research on cosmic rays, since they can be easily sent up to great heights in balloons.

The Wilson chamber records in general the path of any kind of sufficiently energetic charged particles. We shall encounter numerous references to the application of the Wilson chamber to the study of nuclear processes.

Gamma Ray and Alpha Ray Spectra

The length R of the α ray tracks, the so-called *range* (see Fig. 38.6), bears a simple relationship to the initial velocity v_0 of the emerging α particles as was discovered experimentally by Geiger:

$$v_0^3 = cR \qquad (38.1)$$

(where c is a proportionality constant). Hence, measurement of the range of α particles also enables one to determine their initial kinetic energy.

When we examine the length of the tracks of α particles emanating from an α emitter, we find that the recorded tracks can be classified into groups of different ranges ("α ray spectrum") and hence different energies. Figure 38.10 shows schematically two groups of tracks of slightly different energies, E_1 and E_2. If we assume that only one α particle can be emitted by a disintegrating nucleus, we must conclude that in this case the "daughter nuclei" resulting from the emission of an α particle by the "parent nucleus" are not of uniform nature. Those which were formed by emission of α particles of energy E_1 have less energy than the daughter nuclei resulting from the emission of α particles of energy E_2. Does this mean that two different daughter elements have been formed? If so, how do they differ?

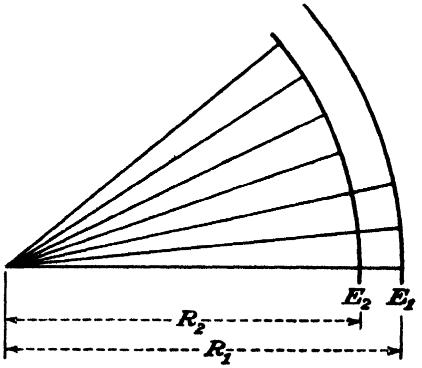


Fig. 38.10.

Their atomic number and mass must be the same! They can differ obviously only by their energy content.

The following observation suggests immediately the answer to the questions posed above. We detect in all cases where groups of α particles of different range are observed the emission of γ rays. In cases such as the above-mentioned example the frequency ν of the emitted γ quanta bears the following relation to the energy of the emitted α rays.

$$h\nu = E_1 - E_2$$

This suggests the following theoretical interpretation:

Nuclei can have excited energy states as well as a stable ground state corresponding to the lowest energy level of the nucleus. When an α particle of the energy E_1 is emitted by the parent nucleus, the daughter nucleus which is formed is created in its ground state. When, on the other hand, α particles of the lesser energy E_2 are emitted, the daughter element is formed in an excited energy state. It possesses an energy surplus of $E_1 - E_2$ over the normal ground state of the nucleus. This energy surplus is emitted (usually within 10^{-12} sec) in form of a photon ($h\nu = E_1 - E_2$) as the daughter nucleus falls into the ground state. Thus we explain γ *radiation as being due to the formation of nuclei in an excited state and subsequent transition to ground level.*

β Ray Spectra

The study of the energy of the β particles which are emitted by some nuclear species presents a striking contrast to the picture of α emission. The momentum of the emitted β rays can be determined by allowing the rays emanating from a point source to move at right angles to a magnetic field. Particles of different momentum are deviated to a different extent. If there were discrete groups of β particles corresponding to different energies, as in the case of the α emission, we should expect to obtain several discrete black spots on a photographic plate upon which the bent β rays are allowed to impinge. We find nothing of that sort. If we discount the electrons that are ejected owing to the photoelectric effect of the γ rays, which often accompany β emission and which form a discrete line spectrum, *we find only a continuous* β *ray spectrum, i.e., we find electrons of all conceivable energy values ranging from zero to a definite maximum energy characteristic of the nuclear species which emits the electrons.*

Does the Law of Conservation of Energy Break Down?

The fact that the parent nucleus emits particles of varying energy suggests again, as in the case of the α particles, that the energy of the resulting daughter nuclei should differ. But in this case we should get an infinite variety of daughter nuclei differing in energy by infinitesimal amounts only. We could hardly assume the resulting nuclei to be in different excited states, since in our past experience with atoms and atomic nuclei such states form a discrete scheme of energy levels whereas here we should have to assume a continuum of states. And besides, we should expect the daughter nuclei to pass eventually into the ground state under emission of a continuous γ ray spectrum, but nothing of that sort is observed!

The assumption that the daughter nuclei formed by β emission are permanently different from each other in energy could not be accepted by physicists. In many different instances such daughter nuclei are radioactive, emitting a discrete spectrum of α and γ rays which indicates clearly that they all behave alike. One could hardly expect nuclei of different energies to show such uniform behavior!

Many leading physicists were led by these observations to doubt the general validity of the law of conservation of energy. They were inclined to assume that the daughter nuclei are all identical in energy content and that the parent nuclei are all in the same initial energy state. Thus the fact that β rays of varying energies are emitted was regarded by a number of scientists as a genuine violation of the law of conservation of energy. It was argued that this law holds statistically but not for an individual process of β emission.

The Neutrino Hypothesis

In order to preserve the validity of the law of conservation of energy Pauli introduced the following hypothesis. He assumed the existence of a tiny neutral particle, the so-called **neutrino,** which is supposed to be emitted simultaneously with a β ray electron. Thus we can assume that all parent nuclei lose the same amount of energy when they are converted into the daughter nucleus, but this energy is now distributed mainly among two particles,† so that only part of this energy is imparted to the electron, the rest being carried away with the neutrino. Thus the continuous electron spectrum is explained by the infinite variety of ways in which the available energy may be distributed among the neutrino and the electron.

So far all attempts to detect the neutrino experimentally have met with great difficulties. Owing to its presumably tiny mass and zero charge, we do not expect it to produce noticeable ionization, so that its direct detection seems almost hopelessly difficult. Evidence accumulated so far is very indirect.

---

† We can neglect the energy imparted to the nucleus because it follows from the laws of conservation of momentum and energy that in an interaction of two masses as different as the electron and the nucleus the kinetic energy imparted to the larger mass is negligible as compared with the kinetic energy imparted to the smaller mass.

Are There Electrons in Nuclei?

The emission of electrons by some radioactive nuclei (such as $_{92}U^{237}$) seems to suggest the existence of electrons in atomic nuclei. But this hypothesis is unacceptable for at least two reasons. (1) The magnitude of the angular momentum and of the magnetic moment of a nucleus cannot be represented as a result of the summation of the magnetic moments and angular momenta of (A) protons and ($A - Z$) electrons. (2) According to the principle of uncertainty there should be an excessively large indefiniteness in momentum of an electron which is confined inside a nucleus of approximate diameter $\simeq 10^{-12}$ cm. This indefiniteness Δp can be obtained from $\Delta p \, \Delta x \simeq h$ by setting $\Delta x = 10^{-12}$ cm. From the resulting indefiniteness in momentum we can estimate the indefiniteness in the energy of the electron. This indefiniteness turns out to be of the order of magnitude of 5×10^7 electron volts, which is about ten times as large as the energy of the fastest electrons seen to emerge in nuclear disintegrations. It is unlikely that electrons of such energy could exist for any appreciable length of time in a nucleus without escaping.

Wave mechanical considerations, which cannot be gone into here, show, however, that particles of electronic charge but of a mass a few hundred times as large as that of an electron can well exist in a nucleus. In order to explain attractive forces between neutrons and protons, Yukawa made the hypothesis that a heavy particle of this kind, to which the name **meson** has been assigned, is being tossed to and fro between a neutron and a proton. When a proton captures the negative meson, it is converted into a neutron, while the neutron which emitted the meson has become a proton. According to this picture it is impossible to say which of the two interacting particles is the neutron and which one the proton, since they constantly change their identity.

Yukawa also predicted that free mesons would have a very short lifetime (in the order of a millionth of a second), being radioactive and decaying into an electron and a neutrino.

Yukawa's predictions were vindicated. One actually observes positive as well as negative particles of a mass intermediate between electrons and protons in the cosmic radiation. But they do not form one single group of particles. Particles of the masses $\sim 215m$, $\sim 282m$, and possibly $\sim 1{,}000m$ have been found (where m is the electronic mass). The radioactivity of the meson has also been demonstrated by catching a meson in the act of giving rise to an electron in a Wilson chamber. Mesons have also been produced in the laboratory by methods which will be mentioned in Chap. 39.

Laws of Radioactive Decay

Let us now survey briefly the sequence of transformations which are exhibited by heavy radioactive elements found in nature. Some of them spontaneously disintegrate by β emission, and others by α emission. The daughter element is usually also radioactive and may emit α or β rays. In some instances a nucleus has a choice of disintegrating by β or by α emission (see the two routes of transition from *RaC* to *RaD* in Fig. 38.11). In all natural radioactive families the *last member of the radioactive family is nonradioactive lead.* The transformations are represented in Fig. 38.11 on a coordinate network so as to show clearly that the emission of a β particle increases the atomic number Z by one unit but does not change the mass number A, whereas the emission of an α particle decreases Z by two and A by four units. Figure 38.11 illustrates the genealogy of a typical radioactive family,† the uranium-radium family. It is interesting to note the transition from UX_1 to UII. It is

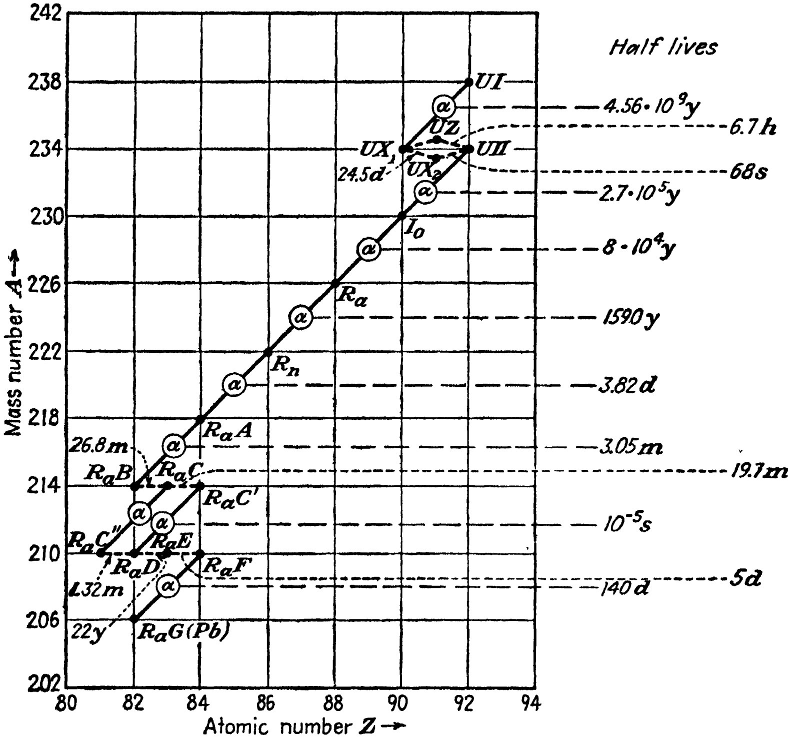


FIG. 38.11. The radioactive uranium-radium family. (The half lives are given in: *y*, years; *d*, days; *h*, hours; *m*, minutes; *s*, seconds.) Beta disintegration is indicated by a horizontal displacement and alpha disintegration by a diagonal (45°) displacement. (The lines UX_1-UZ-UII and UX_1-UX_2-UII are to be imagined as horizontal.)

accomplished by two consecutive β emissions. The first emission can lead to either UZ or UX_2, *i.e.*, to two different elements. Each of these elements decays by β emission, which leads in both cases to UII.

But what is it that makes UZ and UX_2 distinguishable as different elements? Both are formed from UX_1 by β emission, and hence they should have the same atomic number Z and mass number A! Both nuclei decay, leading to the same end product UII! *They differ in a*

---

† Other families are the thorium family and the actinium family.

physical property which is very important in the description of radioactive elements; they have different **half-life periods.** If we start with a population of N UZ atoms, we shall have $N/2$ undisintegrated atoms left 6.7 hr later. If, on the other hand, we start with N UX_2 atoms, half of them will be found to have disintegrated within 1.15 min! Nuclei which have the same A and Z but decay at different rates are called *nuclear isomers.* They can be considered as being the ground state and an excited state of the same nuclear species.

The **half life** of an unstable nuclear species is a very important constant characteristic of the species and descriptive of the "mortality" of its atoms. Its value varies enormously from element to element. We find half lives ranging from fractions of a second to millions of years (see Fig. 38.11 for half lives).

The decay of nuclei is a statistical process analogous to the mortality of human beings under normal living conditions. If we place a sample of a radioactive element in front of a Geiger counter, we hear the clicks which signal a disintegration follow each other in a perfectly irregular haphazard fashion. It appears that an occurring disintegration does not in any way influence the chances of the occurrence of another nuclear disintegration. It proves quite impossible to predict which nucleus will disintegrate next, just as it is impossible to predict at what time a given nucleus will decay. Even though we assume the nuclei of a given unstable element to be perfectly identical, we admit that some of them may "live" a million years before they disintegrate while another identical nucleus may "blow up" a minute from now. *We express our inability to make a "life prognosis" for individual nuclei by stating that the process of radioactive decay is a purely statistical one. We can determine merely what proportion of the original nuclei will disintegrate in a given time but we cannot say which ones will decay.* The situation is similar to the problem of mortality in a given community.

The statistical nature of the alpha as well as of the beta decay can be explained on the basis of quantum mechanics. According to the assumptions of quantum mechanics a relation of the form $\Delta p_x \, \Delta x \gtrsim h$ not only exists between the uncertainties in position and momentum but applies to all pairs of physical quantities for which the product of their dimensions has the dimension of an action, namely, (force × distance × time). Energy and time fulfill this condition, so that the following uncertainty relation should hold for simultaneous determination of energy and time:

$$\Delta E \, \Delta t \gtrsim h$$

Now in the radioactive decay processes the energy liberated in a nuclear disintegration is very sharply defined. In other words, the variation ΔE in the kinetic energy of emitted α particles is very minute, and we also assume in the β decay the total kinetic energy of the electron plus neutrino to be very nearly the same for all disintegrating nuclei, except for a *very small* indefiniteness ΔE. Owing to the extreme smallness of ΔE the indefiniteness Δt in the lifetime of a nucleus becomes enormously large. This accounts for the fact that some nuclei may "live" millions of times as long as others and explains why no definite "length of life" can be stated for a given nucleus.

We can express the rate of "nuclear mortality" as follows: Assume that you have N nuclei of which ΔN disintegrate on the average during the short time interval Δt. Then the fraction of the total number of nuclei which will decay on the average in Δt seconds is $\Delta N/N$, and the fraction of the nuclei decaying per second on the average, *i.e.*, **the rate of decay,** is

$$\boxed{\frac{\Delta N/N}{\Delta t} = -\lambda} \qquad (38.2)$$

where λ (which is not to be mistaken for a wave length!) *is the so-called* **decay constant.*** *It measures the rate of decay of a nuclear species.* It can be shown mathematically that this equation is equivalent to†

$$\boxed{\ln\left(\frac{N_0}{N}\right) = \lambda t} \qquad (38.3)$$

where N_0 is the initial number of nuclei and N the undisintegrated residue after t seconds.

There is a simple relationship between the decay constant λ and the "half life" T to which we referred above: We apply equation (38.3) to the case where $N = N_0/2$, *i.e.*, we wait $t = T$ seconds until half of the originally present nuclei have disintegrated. Then we get

$$\ln\left(\frac{N_0}{N_0/2}\right) = \lambda T$$

or

$$\ln (2) = \lambda T = 0.693$$

hence,

$$\boxed{T = \frac{0.693}{\lambda}} \qquad (38.4)$$

Thus we can easily compute λ if we measure T or vice versa.

---

* The minus sign indicates that N decreases as t increases. [Compare equation (34.1*a*), $\frac{\Delta N}{N} \Big/ \Delta x = -\mu$.]

† ln (N_0/N) is the natural logarithm of N_0/N; *i.e.*, it is the power to which the number $e = 2.718 \ldots$ must be raised in order to obtain N_0/N. ln (N_0/N) is defined by the equation $N_0/N = e^{\ln(N_0/N)}$.

For elements with a small value of λ the decay constant can be determined by measuring with a Geiger counter how many disintegrations (ΔN) occur during Δt seconds. If N, the number of the atoms in the original disintegrating sample, is known, λ is obtained from equation (38.2) and hence also T from equation (38.4).

Geiger and Nuttal discovered empirically a very important relation which exists between the rate of decay and the velocity (and hence also energy) of emitted α particles in the case of α-emitting nuclei:

$$\boxed{\log \lambda = A + B \log v} \tag{38.5}$$

where λ is the decay constant, v the velocity of the emitted α particles, and A and B are constants. This equation allows us to determine the energy of the emitted α particles by measuring the half life of the element or vice versa. This equation is of additional interest to us, since it has been possible to derive it theoretically (see page 794). The success of deriving it may be considered as a strong support for the highly interesting hypothesis on which the derivation is based.

The following example shows the enormous variation in the so-called *mean life* (defined as $L = 1/\lambda$) of an α emitter with the energy of the emitted α particles:

Element	$\frac{1}{\lambda}$ (mean life)	Energy, electron volts
Th	2×10^{10} years	4.00×10^6
RaC′	2×10^{-4} sec	7.68×10^6

We see that an increase of energy merely by a factor of about 2 decreases the mean life by a factor of about 10^{20}!

Quantum Mechanical Theory of α Decay

How can this enormous variation of the mean life (and hence of the decay constant λ) with the energy of the emitted α particle be explained? Classical physics proved completely incapable of answering this question. A satisfactory explanation was finally provided by Gamow (and independently by Condon and Gurney) on the basis of wave mechanics. In order to understand their theory let us begin with the following preliminary consideration: Let us assume that we move a positive point charge q toward a stationary positive point charge Q. We are doing work in this process. The potential energy of q increases according to the function $V = Qq/r$, which follows from Coulomb's law. This expression holds, as we have seen, for the approach of an α particle toward the nucleus in Rutherford's scattering experiments. The Coulomb potential energy of the α particle as a function of its distance from the center of the nucleus looks as shown in Fig. 38.12 by the solid curve.

This situation is analogous to the case of a hill of the given shape upon which a ball is allowed to roll. As the ball rises while moving to the left, its potential energy increases. When it has reached the height at which its potential energy is equal to its original kinetic energy, it begins to roll back down to the right. In describing the nuclear field we shall henceforth speak of a potential hill or a potential barrier in the sense of the above analogy. We see that the downward slope of the curve to the right (as a result of which the sphere in our analogy is made to move to the right) indicates forces of repulsion from the nucleus. The nucleus is assumed to be at the origin of the coordinate system at O.

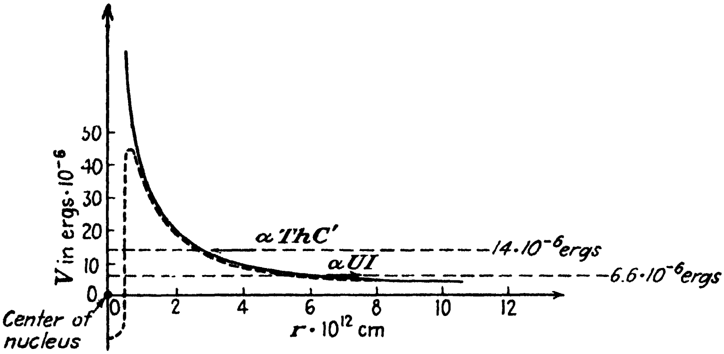


FIG. 38.12. Potential energy of an α particle as a function of the distance from the center of the nucleus.

Rutherford found from his scattering experiments with α particles that Coulomb's law holds well down to distances of the order of magnitude of 10^{-12} cm. The deviations from this law which are observed for closer approach of the α particles to the nucleus are explained by attractive forces between the nucleus and the α particle which become effective at close range. The potential energy of an approaching α particle ceases to increase at the distance at which the force of repulsion becomes equal to the force of attraction (highest point of the dotted curve in Fig. 38.12) and decreases under the action of predominantly attractive forces as we come closer to the center of the nucleus. Thus the resultant potential function of the nucleus is represented by the dotted curve rather than by the solid curve which represents the potential function due to the Coulomb forces only. If we imagine this curve rotated about the vertical axis, we obtain a round wall resembling the shape of a volcano with a deep crater.

If an α particle were to move fast enough toward the nucleus, it could penetrate into the "potential crater," just as a sufficiently fast ball would roll over the top of the potential hill into its crater. To do this the α particle (or the ball) would have to have an initial kinetic energy at least equal to the potential energy it would possess at the top of the hill (since this potential energy must be acquired at the expense of the kinetic energy of the particle).

The lower limit for the height of the "potential crater" has been estimated as follows: It has been shown that α particles of ThC' of energy 14×10^{-6} erg are scattered backward from nuclei of UI in accordance with the Rutherford formula. This means (see Chap. 35 on estimate of the nuclear radius) that the scattering is entirely due to Coulomb repulsion. In terms of the Fig. 38.12 it means that the particles do not reach the top of the potential crater where the potential function begins to depart appreciably from the Coulomb potential. Since the highest point on the potential curve reached by the α particle is given by its kinetic energy of 14×10^{-6} erg, *we conclude that the top of the potential crater corresponds to a higher energy than this* (see horizontal dotted line labeled $\alpha ThC'$ in Fig. 38.12).

Uranium I (U^{238}) disintegrates, emitting α particles. According to our rolling-ball analogy, we can imagine α particles to move about inside the crater with a certain velocity which should enable them to reach the top of the hill and "roll" over it and out of the crater. The smallest kinetic energy we should expect these α particles to have at infinite distance from the nucleus would be equal to their potential energy at the top of the hill, which is converted into kinetic energy as they "roll" down the hill. We just saw that the potential energy of the α particles at the top of the hill must be greater than 14×10^{-6} erg, since this is the energy of the ThC' α particles which have not been able to reach the top. Thus, we expect the energy of the α particles emitted by uranium I to be at least 14×10^{-6} erg, since they must have "rolled down the potential hill." Surprisingly enough, it is not! It is only 6.6×10^{-6} erg!! (See horizontal line labeled *UI* in Fig. 38.12.)

How could they get out of the crater? Gamow, and Condon and Gurney explained this by the wave mechanical "tunnel effect," or penetration of potential barriers, discussed in Chap. 37.

According to this theory a moving particle inside the potential barrier has a finite chance to "leak" out of the barrier even if its kinetic energy is less than required by classical physics to carry it over the potential barrier. The probability of leaking out, *i.e.*, of emission of a particle, increases very rapidly as the kinetic energy of the corpuscles is increased. The particles are imagined to move inside the potential crater, running repeatedly against the potential barrier, the picture being similar to that of gas molecules in a container. The theory predicts that a nuclear species containing faster particles inside the nuclear potential crater will be losing them at a faster rate, owing to tunnel effect and to more frequent impact of particles against the potential barrier, than a nuclear species whose crater is populated by slower particles. Thus a correlation is to be expected between rate of emission of α particles and the speed of the emitted particles. Such a correlation between the decay constant λ and speed of emitted α particles had been previously established by experiments [Geiger-Nuttall law, equation (38.5)]. It has actually proved possible to derive on the basis of the tunnel-effect theory a relationship between λ and the energy of the emitted α particles which is equivalent to the Geiger-Nuttall law. This was the first successful application of wave mechanics to nuclear theory.

APPENDIX

Radioactivity and Temperature of the Earth

If we insert a pipe deep into the ground and pump air from the depth to the earth's surface, we find it to be radioactive owing to the presence of radioactive gases such as radon. Chemical analysis of the soil explains the presence of these emanations. One finds radioactive elements such as uranium, radium, and thorium in the ground which give rise to gaseous radioactive decay products. The radioactive decay of these elements liberates a considerable amount of heat. This is an important observation, which enables us to solve a puzzle which worried physicists before the discovery of radioactivity. The amount of energy which the earth receives by radiation from the sun is insufficient to compensate for its radiative heat losses. As a result, the surface of the earth should be getting cooler. We know, however, from geological evidence, that the surface temperature of the earth must have varied very little during the last few hundred million years, having remained in the temperature range which permits the existence of life. How can we account for this approximate constancy of the earth's temperature? The radioactive decay of the uranium and thorium content of the earth is more than ample to provide the additional amount of energy required to maintain the surface of the earth at constant temperature. We do not know how far below the earth's surface the uranium and thorium content extends, but it is sufficient to make the conservative estimate that only 2×10^{20} gm of thorium and 4.5×10^{20} gm of uranium are contained in a relatively thin superficial layer. The radioactivity of the slightly radioactive potassium must also be taken into consideration. One has estimated that the beta decay of potassium possibly contributes as much energy as the combined radioactivities of thorium and uranium.

The Age of the Earth

We have seen in this chapter that stable lead is the end product of all radioactive family trees. If we imagine crystals of uranium compounds to have been formed during the period of solidification of the earth's crust, we can easily see how these crystals could be used as a clock to measure the time which elapsed since that crystallization took place. The decomposition of uranium leads to lead, which must remain embedded in the crystal of the uranium compound. Thus knowing the decay constant of uranium and measuring the amount of lead found in the crystal we can compute how long a time was required to produce that amount of lead. Since the earth must be older than the uranium compound crystals, this method allows us to determine the lower limit for the age of the earth. The accepted estimate obtained by this method is about 2×10^9 years.

Instead of measuring the lead content one can also measure the helium trapped in the mineral. Each helium atom represents a previously emitted alpha particle. In making such estimates one must remember that each uranium I atom, for instance, gives rise to eight alpha particles which are emitted consecutively from the various daughter elements of uranium I.

The helium accumulation method has been also used to estimate the age of meteorites, which should give us a lower limit for the age of the universe. Thus, Paneth, Urry, Evans, and Goodman have succeeded in the very difficult task of measuring the helium content of meteorites which amounts to a fraction of a cubic centimeter of gas per ton of rock. The oldest meteorites they found did not exceed the age of 2.8×10^9 years.

QUESTIONS AND PROBLEMS†

1. How could you determine practically the decay constant of a radioactive element (*a*) of long half life? ($\sim 10^6$ years) (*b*) of short half life (~ 5 hr)?

2. What is the mechanism of heat generation in a sample of radium?

3. How does the character of the radiation of a freshly prepared sample of radium change in the course of time?

4. What conclusions can you draw about the interactions of gamma rays with electrons from the fact that they penetrate matter with little absorption?

5. How do we distinguish an electron from an alpha particle in a cloud-chamber photograph?

6. We know that the ionization density (*i.e.*, number of ion pairs produced per centimeter) increases near the end of an alpha particle track. How do you explain it?

7. A nucleus of species A can decay into a nucleus of species B in one of two ways. (*a*) It can emit an alpha particle of 2×10^6 electron volts and form nucleus B, or (*b*) it can emit an alpha particle of 1.8×10^6 electron volts, then a gamma ray, the result of which is again a stable nucleus B. What is the wave length of the emitted gamma ray quantum?

8. A nucleus of species A emits alpha particles of energy 4×10^6 electron volts, while a nucleus of species B emits alpha particles of 2.5×10^6 electron-volt energy. Which of the two elements has the longer half life?

9. How does Gamow's theory explain the fact that one nucleus of radium may decay by alpha emission 10 min from now whereas another identical radium nucleus may live another 1,000 years before decaying into radon?

10. Find the loss of mass of a nucleus which emits a γ ray quantum of $\lambda = 0.01$ A.

11. Imagine two oppositely charged points (1 coulomb each) 0.01 mm apart. Determine the change in mass of this system when the two charges are moved apart to a distance of 10 m. Is it a gain or a loss of mass?

12. A potassium atom emits a β ray electron *of maximum velocity*. What is the ratio of the kinetic energies between the emitted electron and the recoil nucleus? Would you be able to answer the question if the electron were not one of maximum velocity?

---

† See Table I, p. 867.

13. An element emits in succession three α particles and two electrons.

a. What is the change in its atomic number?

b. What is the change in its mass number?

c. How much does its mass change (1) as expressed in grams, (2) in amu? (Neglect changes in mass due to conversion of mass into energy.)

14. A nucleus emits a beta ray electron of 500,000 electron volts. If the energy of the fastest electrons emitted by this nuclear species is 800,000 electron volts, what energy would you ascribe to the neutrino emitted in this case?

15. Get the values for several different half lives from Fig. 38.11 and compute the corresponding decay constants.

16. *a.* What fraction of radium atoms disintegrate each second? (1 year = 3.15×10^7 sec.)

b. How many alpha particles are emitted in a second from 1 mg of pure radium?

c. How long would it take for 10^{-3} per cent of atoms to disintegrate in a sample of pure radium? The half life of radium is 1,590 years.

17. A polonium nucleus emits an α particle of 5.3×10^6 electron volts. Compute the kinetic energy of the resulting lead nucleus.

CHAPTER 39

ARTIFICIAL TRANSFORMATION, CREATION, AND ANNIHILATION OF THE BUILDING BLOCKS OF MATTER

a. Is the statement that in all radioactive transmutations by bombardment with particles an intermediate unstable nucleus is formed a statement of fact or a hypothesis?

b. What observation led to the discovery of artificial transmutation of elements?

c. How can the energy of the emitted nuclear fragments be measured?

d. How can the validity of the laws of conservation of (1) momentum, (2) energy, (3) mass be tested in nuclear transformations?

e. Are nuclear transmutations "endothermal" or "exothermal"?

f. What is the source of energy in "exothermal" nuclear transformations?

g. What projectiles can be used for nuclear transmutation?

h. Which do you consider the most efficient projectiles for nuclear transmutation?

i. Name some sources of high-energy particles. Name some sources of neutrons.

j. What properties do the neutrons have?

k. How can neutrons be detected?

l. Do we know whether neutrons are elementary particles or a close combination of a proton with an electron?

m. Do we know whether protons are elementary particles or a close combination of a neutron with a positron?

n. How can neutrons be slowed down?

o. How was it shown that neutrons are particles of nearly the same mass as the proton rather than a high-energy gamma radiation?

p. What happens eventually to a neutron (1) in matter? (2) in a vacuum?

q. What is a positron?

r. What happens to a positron eventually?

s. How can positrons be produced?

t. What is the meaning of the term (1) "pair creation"? (2) "annihilation" of positrons and electrons?

u. Give an example of a direct transformation of (1) material particles into energy, (2) energy into material particles.

v. What is a positron according to Dirac's theory?

w. What process is designated as "induced radioactivity"? Name an example of this process.

x. What particles are usually emitted by elements which have been rendered artificially radioactive?

y. What "projectile" is most effective in yielding artificially radioactive nuclei?

z. How are "transuranium" elements formed?

aa. Name at least three characteristics of nuclear fission which distinguish it from other modes of nuclear transmutation.

bb. Which feature of nuclear fission makes a chain reaction possible?

cc. How does the liquid-drop model of the nucleus "explain" fission?

dd. How does a cyclotron work?

ee. What is the principle of a uranium pile?

ff. What are the uses of artificial radioactive elements in biological and medical research?

Can the Process of Radioactive Decay Be Reversed?

In Chap. 38 we considered a hypothesis of alpha disintegration of nuclei. We assumed that at least some of the nucleons are grouped into the complex of two neutrons plus two protons, thus forming an alpha particle. (It is immaterial whether such grouping is permanent or temporary and recurring.) The alpha particles were assumed to be moving inside the nucleus with considerable kinetic energy, repeatedly bouncing off from the "walls of the potential barrier" which we may consider as the boundary of the nucleus. According to the wave mechanical theory an alpha particle may succeed, after a number of "unsuccessful trials," in "rolling over" or "leaking through" the wall of the potential crater and thus escape with a considerable kinetic energy.

But why should this process occur only in a small group of the heaviest elements? Are the nuclei of the so-called *nonradioactive elements* of an essentially different constitution? It is not necessary to assume that. It may well be that many of the heavier so-called nonradioactive elements are actually radioactive but have decay constants so small that it is hopeless with the experimental means at our disposal to detect the exceedingly infrequent decompositions of their nuclei.

Having seen how new nuclear species are formed by the escape of alpha particles from the interior of nuclei, *we may ask the question whether new elements can also be created by the converse process of allowing alpha particles to "roll over the wall of the potential crater" into the nucleus.* The same theory which showed that an alpha particle should be able to get out of the nucleus even if its kinetic energy is too small for an escape, according to classical physics, predicts that the converse is also true; namely, an alpha particle has a finite chance of entering a nucleus even if its kinetic energy is insufficient, according to classical physics, for penetration of the nuclear potential barrier. The probability of penetration increases very rapidly with the kinetic energy of the alpha particles. Thus nuclear synthesis appears to be at least a theoretical possibility.

Artificial Transmutation of Atoms

Nuclear synthesis was eventually demonstrated experimentally but not as an outcome of the above-mentioned considerations. It grew out of a discovery which was an accidental by-product of scattering experiments, which, in various modifications, have contributed so much to our knowledge about the atoms. In 1919 Rutherford was investigating scattering of alpha particles which were passing through an atmosphere of nitrogen.† Such experiments can be made with Rutherford's scattering apparatus of the type shown in Fig. 35.4 (the foil being removed so as to use the intervening gas as scatterer) where the scattered particles are detected by the scintillations they produce on a screen or, more clearly, by means of a Wilson cloud chamber which indicates the tracks of the individual particles involved in the scattering process. Figure 39.1 is a typical Wilson chamber photograph showing a collision between an alpha particle and an oxygen atom, while Fig. 39.2 shows an alpha collision with a hydrogen atom. The track of the recoiling proton (Fig. 39.2) is seen to be thinner and longer than the track of the recoiling oxygen nucleus (Fig. 39.1). Pictures of collision of α particles with nitrogen nuclei are very similar to those of collisions with nuclei of oxygen.

Fig. 39.1. Collision of an α particle with an oxygen nucleus (see forked track). The shorter track is due to the recoiling oxygen atom. (*Courtesy of P. M. S. Blackett and Proceedings of the Royal Society.*)

For a given angular deviation of the alpha particle from its original path and for a given direction of motion of the recoiling nucleus with which it collides, it is possible to compute from the laws of conservation of energy and momentum the energy of the recoiling nucleus and therefrom its range, making use of a relation between range and particle speed which has been found experimentally. Measurements of the tracks observed in the Wilson chamber are in excellent agreement with such predictions. Thus, for instance, alpha particles of range $R_\alpha = 7$ cm moving through hydrogen are expected to produce recoil protons up to the range of $R_p = 28$ cm. Rutherford noticed, however, when shooting alpha particles of $R_\alpha = 7$ cm through *nitrogen*, the occurrence of particles of the unexpectedly long range of $R = 40$ cm. Determination of e/m of these particles by bending them in a magnetic field showed them to be protons. These protons presented a puzzle, since no trace of hydrogen was present in the gas. But even if it were present the range of the recoil protons should not have exceeded 28 cm.

---

† Rutherford was using a scintillating screen, but for greater clarity, we shall discuss the experiments as they can be performed with a Wilson chamber.

How can this observation be explained? Rutherford suggested a hypothesis which we quote in an extended version given it by Bohr: *Fast-moving nuclei are enabled by their great kinetic energy to approach each other very closely in a collision process. At such distances nuclear attractive forces come into play which bind the two nuclei together to form a new nucleus whose atomic number is equal to the sum of the atomic numbers of the colliding nuclei and whose mass number is equal to the sum of the mass numbers of the nuclei. The newly formed nucleus is unstable and undergoes a transformation into a stable one by emission of a proton.*

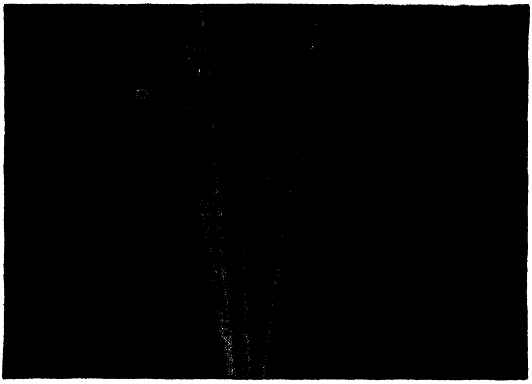


FIG. 39.2. Collision of an α particle with a hydrogen nucleus (see forked track). The light long track is that of the hydrogen nucleus. (*From Rutherford, Chadwick, and Ellis, "Radiations from Radioactive Substances," Cambridge.*)

In the above example we assume that the helium nucleus (alpha particle) combines with the nucleus of nitrogen to form an unstable nucleus of fluorine which immediately disintegrates into oxygen and hydrogen (emitted proton) according to the following reaction:

$$_2He^4 + {}_7N^{14} \rightarrow ({}_9F^{18}) \rightarrow {}_8O^{17} + {}_1H^1 \tag{39.1}$$

(39.1) cannot be interpreted as an equation, since the sum of the masses on the right side is larger than the sum of the masses on the left side.

We can, however, convert it into an equation by taking into consideration the energies of the reacting nuclei which can be expressed in terms of mass according to Einstein's law: $m = E/c^2$. In the following equation the chemical symbols stand for the nuclear masses† in amu (1 amu = one-sixteenth of the mass of an O^{16} atom), A is the kinetic energy of the alpha particle in amu, and B is the energy of the emitted proton in amu. (Why are we justified in neglecting the energies of the other participating particles?) The experimentally determined data listed below the symbolic equation prove the validity of Einstein's mass-energy law:

$$\begin{array}{c} {}_2\mathrm{He}^4 + {}_7\mathrm{N}^{14} + A = {}_8\mathrm{O}^{17} + {}_1\mathrm{H}^1 + B \\ \underbrace{4.0039 + 14.0075 + 0.0083}_{18.0197} = \underbrace{17.0045 + 1.0081 + 0.0071}_{18.0197} \end{array} \qquad (39.2)$$

This is a so-called *endothermic reaction; i.e.*, the amount of energy liberated is smaller than the amount of energy invested ($B < A$).

Rutherford's discovery of artificial disintegration of nitrogen led immediately to an avalanche of investigations by numerous workers who found similar transmutations of other elements. Thus, for instance, it was found that boron or aluminum bombarded by sufficiently energetic alpha particles is transmuted into carbon and silicon, respectively, under emission of protons according to the following reactions:

$${}_5\mathrm{B}^{11} + {}_2\mathrm{He}^4 \rightarrow {}_6\mathrm{C}^{14} + {}_1\mathrm{H}^1 \qquad (39.3)$$

$${}_{13}\mathrm{Al}^{27} + {}_2\mathrm{He}^4 \rightarrow {}_{14}\mathrm{Si}^{30} + {}_1\mathrm{H}^1 \qquad (39.4)$$

Examination of the kinetic energies of the particles involved reveals that the latter reaction is *exothermal.* The sum of the masses on the left side of the arrow is larger than on the right side. As a result the energy of the emitted proton is larger than the energy of the bombarding α particle. *This process could be used (at least in principle) as a source of energy.*

Rutherford's discovery had a double significance: (1) It showed the possibility of transmutation of elements, which was for many centuries the dream of alchemists, and suggested that Prout may have been right in assuming all atoms to be made up of the same kind of "primordial matter." (2) It indicated a new way of obtaining power which is much more productive of energy per reacting particle than any chemical reaction we know.

---

† The masses used in the numerical example are atomic masses rather than nuclear masses; *i.e.*, they include the mass of the planetary electrons. This is permissible because the number (and hence the mass) of the electrons is the same on both sides of the equation; *i.e.*, we commit the same "error" on both sides of the equation.

Transmutation by Artificially Accelerated Particles

The sources of alpha particles used by Rutherford and subsequent investigators consisted of samples of radioactive alpha emitters. Such sources are not very satisfactory for research purposes because (1) they are usually very weak (*i.e.,* relatively few particles are emitted per second); (2) the emitted alpha particles possess a definite speed whereas it would be of great interest to study nuclear reactions over a range of continuously varying speed; (3) theory leads us to expect (see page 794) that the number of successful penetrations of the nuclear potential barrier would greatly increase with the speed of the alpha particles, so that it would be of advantage to use particles of greater speed than those available from the naturally radioactive sources; and finally (4) alpha particles are not the most favorable projectiles for nuclear bombardment. When a positively charged particle approaches a nucleus, it is slowed down by the coulomb repulsion and thus loses kinetic energy. Obviously a particle, such as a proton, carrying only a single positive charge will come closer to a nucleus than a doubly positively charged projectile, such as an alpha particle, of equal kinetic energy.

Such considerations stimulated the development of intense sources of protons (or deuterons, nuclei of $_1H^2$) and special accelerators in some of which the hydrogen ions could be accelerated to attain energies in excess of the energies of the naturally available alpha particles. The simplest scheme utilizes an electrostatic high-voltage generator (van de Graff machine) in which the ions are accelerated through a potential difference of a few million volts. But insulation difficulties at high voltage make it very difficult to advance far above several million volts of accelerating voltage. Ingenious schemes, such as the *cyclotron* (see Appendix), have been suggested which enable us to accelerate ions to energies of hundreds of millions of electron volts without employing potential differences exceeding some hundred thousand volts.

The first successful transmutation of nuclei by artificially accelerated protons was carried out in Rutherford's laboratory by Cockroft and Walton. They were constructing a transformer generator of high voltage for proton acceleration. At the request of Lord Rutherford, who did not have the patience to wait until the completion of the installation, they performed a bombardment experiment at a stage of development when their generator yielded an accelerating voltage of only 150,000 volts. They chose as their target the lightest known solid, lithium.

Since they were not thinking in terms of the "tunnel effect," they were surprised to find that nuclear transmutation actually was produced by 150,000-electron volt protons. In fact, soon afterward Rausch von Traubenberg showed that the observed transmutation of lithium took place even with as low an accelerating voltage as 13,000 volts! The only effect of the accelerating voltage was shown to be in determining the efficiency of the process. Not every proton transmutes a lithium nucleus. At low proton energies the percentage of "successful" protons (*i.e.*, of those which cause a transmutation) is vanishingly small, but it rises very rapidly with increase in the kinetic energy of the protons. But even at 150,000 volts only a very small fraction of the bombarding protons is effective in causing a nuclear reaction. Most of the kinetic energy of the protons is transformed into heat, raising the temperature of the target. The nuclear reaction observed by Cockroft and Walton was as follows:

$$\underset{\substack{\uparrow\\ \text{proton}}}{{}_1\text{H}^1} + \underset{\substack{\uparrow\\ \text{target}}}{{}_3\text{Li}^7} + \underset{\substack{\uparrow\\ \text{energy}\\ \text{of the}\\ \text{proton}\\ \text{in amu}}}{E_p} = ({}_4\text{Be}^8) = \underset{\substack{\uparrow \qquad\quad \uparrow\\ \alpha\text{ particle}}}{{}_2\text{He}^4 + {}_2\text{He}^4} + \underset{\substack{\uparrow\\ \text{energy}\\ \text{of each}\\ \alpha\text{ particle}\\ \text{in amu}}}{2E_\alpha} \qquad (39.5)$$

The ${}_4\text{Be}^8$ nucleus was not observed; it is merely written for convenience as a hypothetical intermediate step. We assume that the proton combines with the lithium nucleus to form an unstable ${}_4\text{Be}^8$ nucleus of exceedingly short half life which disintegrates practically instantaneously into two alpha particles (${}_2\text{He}^4$) of equal energy.

Range measurements reveal that the energy of each of the resulting alpha particles in the above reaction is roughly 8.5×10^6 electron volts. Thus by using only 150,000 electron volts of energy to accelerate the proton, we gain about 17×10^6 electron volts as kinetic energy of two alpha particles, or more than 100 times as much energy as we invested! This process looks like a very lucrative source of energy. But unfortunately, this impression is due to an oversight! As we mentioned above, a very small percentage of the bombarding protons is effective in transmuting a lithium nucleus, so that actually the energy "yield" from transmutations is very small. Most of the energy we invest is wasted, being transformed into heat; the energy liberated in the process of transmutation is only a small fraction of the energy input.

We can also compute the kinetic energy of the two alpha particles to be expected and we can compare the theoretical value with the experimental value of 17.28×10^6 electron volts. In equation (39.6) the energy of the 150,000-volt protons is expressed in amu as computed from Einstein's law $m = E/c^2$.[†]

$$\begin{array}{lccccccc} & {}_1\mathrm{H}^1 & + & {}_3\mathrm{Li}^7 & + & \mathrm{E}_p & = 2({}_2\mathrm{He}^4) + 2(E_\alpha) & (39.6) \\ \text{Masses in amu:} & 1.0081 & + & 7.0182 & + & 0.00016 & = 8.0078 + 2(E_\alpha) & \end{array}$$

From this follows $2E_\alpha = 0.0187$ amu $= 17.4 \times 10^6$ electron volts, which is in good agreement with the approximate experimental value of 17.28×10^6 electron volts.

The study of many nuclear transformations has been carried out with great precision as to the amount of energy released and the loss of mass. Such studies have fully confirmed Einstein's mass-energy law.

DISCOVERY OF NEW PARTICLES

Neutrons

The study of the bombardment of various nuclei by fast-moving alpha particles led not only to the discovery of nuclear transmutations accompanied by ejection of protons but also to ejection of new elementary particles. In 1930 Bothe and Becker observed that a very penetrating radiation, capable of traversing several feet of lead (!), emanated from a beryllium target while the latter was bombarded by alpha particles from a polonium source. They assumed this radiation to consist of gamma rays of unusually high penetrating power.

Their experiments were repeated by Irène and Frédéric Joliot-Curie, who made a very strange observation. Usually when we interpose a screen of matter between a source of radiation and a suitable detector, the intensity of radiation recorded by the detector is diminished owing to absorption of radiation by the interposed screen. The Joliot-Curies found, however, that the reading of an ionization chamber (*C* in Fig. 39.3) is greatly *increased* by interposing a paraffin block *P* between the beryllium target *Be* which is subjected to bombardment by alpha particles from the polonium source *S* and acts as the source of the mysterious, penetrating radiation discovered by Bothe and Becker.

The Joliot-Curies soon established that the increase in ionization in the chamber *C* was caused by ejection of protons from the paraffin

---

[†] 1 amu $= 1.66 \times 10^{-24}$ gm; 1 electron volt $\approx 1.60 \times 10^{-12}$ erg. Hence from $m = E/c^2$,

$$\underline{\text{1 electron volt}} = \frac{1.60 \times 10^{-12}}{(3 \times 10^{10})^2} = 1.78 \times 10^{-33}\ \text{gm} = \frac{1.78 \times 10^{-33}}{1.66 \times 10^{-24}}\ \text{amu}$$

$$= \underline{1.07 \times 10^{-9}\ \text{amu}}$$

For 150,000-electron-volt protons the mass equivalent of energy is:

$$1.5 \times 10^5 \times (1.07 \times 10^{-9}) = 1.60 \times 10^{-4}\ \text{amu}.$$

plate P by the penetrating radiation. To explain the ejection of protons they formulated the following hypothesis: Alpha particles of polonium excite beryllium nuclei to emit an exceedingly penetrating gamma radiation of an extremely short wave length. These gamma ray photons of enormously large energy collide elastically with the hydrogen nuclei contained in paraffin† and eject them by collision in a similar fashion as "recoil electrons" are ejected during the process of Compton scattering of gamma rays and X rays.

The protons ejected from the paraffin block had a range of 40 cm in air, which corresponds to a proton energy of 5.7×10^6 electron volts. Chadwick calculated that, in order to be able to impart such a large energy to a proton in a head-on collision, a gamma quantum should possess an energy of at least 55×10^6 electron volts. The emission of γ ray quanta of such large energy from the beryllium target bombarded by α particles seemed inexplicable. But the main difficulty arose when the following contradictory result was obtained when the penetrating radiation was allowed to fall upon a target containing nitrogen atoms. It was observed that nitrogen ions of 1.2×10^6 electron volts are knocked out of the nitrogenous material. To impart this energy to a nitrogen nucleus in a head-on collision a γ-ray quantum would have to possess an energy of 90×10^6 electron volts. Thus experiments with the same γ quanta lead to widely different estimates of the energy of the quanta using different methods of observation.

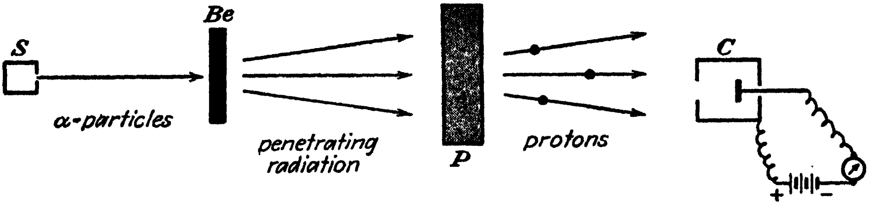


FIG. 39.3.

Chadwick concluded from these results that the photon hypothesis is untenable and showed that *the contradiction disappears if one assumes the penetrating radiation to consist of neutral material particles of a mass approximately equal to the mass of a proton. He was not unprepared for this discovery.* He was well aware of the fact that his teacher Rutherford had been searching for a decade for a particle whose existence he suspected and which he called a **neutron.** Rutherford assumed the possible existence of particles resulting from a closer connection between a proton and an electron than that which exists in a hydrogen atom. He expected that such neutral particles would be hard to detect because, since they do not carry an unbalanced electric charge, they would not ionize matter. Nevertheless such particles could interact with atomic nuclei in close encounters. In a central elastic collision with a particle of equal mass a moving particle stops and communicates its entire momentum to the stationary particle, which then continues moving with a momentum equal to that of the particle which hit it. This behavior is a simple consequence of the laws of conservation of momentum and energy in an elastic impact. This is the approximate mechanism through which the proton in the paraffin can be imagined to acquire momentum in collisions with neutrons.

---

† Paraffin can be replaced by other compounds containing hydrogen atoms.

Chadwick was able to determine the mass of the neutron from cloud-chamber studies of collisions of neutrons with different nuclei. From the laws of conservation of momentum and energy he was able to derive the following simple equation for head-on collisions:

$$v = \frac{2m_n}{m + m_n} v_n \qquad (39.7)$$

where v_n and m_n are the initial velocity and mass of a neutron and v and m the velocity and mass of a nucleus which collides with a neutron. The value of v can be determined by measuring in a cloud chamber the range of the nucleus which has been hit by a neutron. Thus we can observe the range of recoil nuclei of hydrogen by allowing neutrons to enter a Wilson chamber filled with hydrogen. Figure 39.4 shows a Wilson chamber photograph of protons most frequently knocked nearly in the forward direction by neutrons coming from a source at the lower right corner. Since we know the mass m of the proton and can determine the velocity v of the protons from their range, we obtain by substituting these data in equation (39.7) an equation with the two unknowns v_n and m_n (the velocity and the mass of the neutrons). By performing the same type of observation with neutrons issuing from the source and entering an atmosphere of nitrogen we obtain a second equation with the same unknowns m_n and v_n by substituting in equation (39.7) the values of the mass and the observed velocity of the recoil nitrogen nuclei. From the two resulting equations, we can compute m_n, the **mass of the neutron.** Chadwick found†

$$\boxed{m_n = 1.00890 \text{ amu}} \qquad (39.8)$$

---

† This value is based on a more exact method of determination by Chadwick and Goldhaber.

The mass of the neutron is seen to exceed the mass of the proton $m_p = 1.00812$ by about $1\frac{1}{2}$ electron masses (the mass of the electron being $m \approx 5.5 \times 10^{-4}$ amu).

Thus the neutral elementary particle which we assumed to be one of the two building blocks of matter (all nuclei being assumed to be built up of neutrons and protons) was actually isolated experimentally. We shall consider neutrons to be nuclei of the first element in the periodic table, an element of atomic number $Z = 0$ and mass number $A = 1$. Since the nuclei of this element are neutral, there is no orbital electron. Hence, this element is invisible, having neither an emission nor an absorption spectrum.

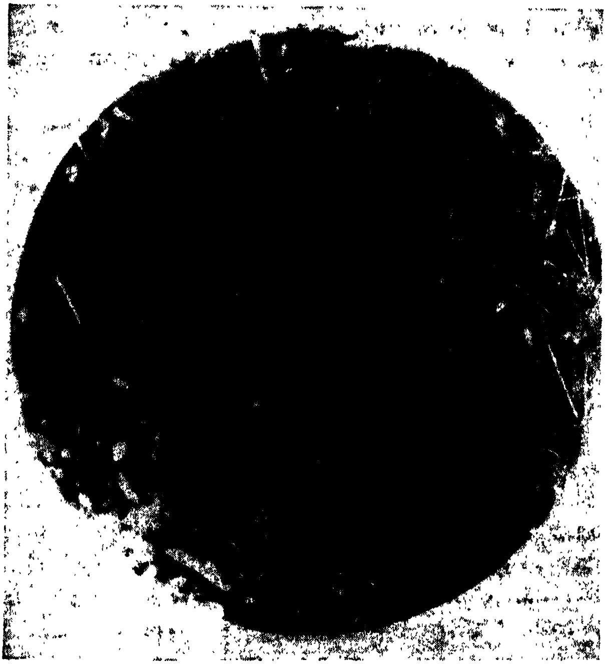


FIG. 39.4. Tracks of protons which have acquired great momentum by collision with fast neutrons. The neutrons (no visible tracks) are moving in the direction from the lower right to the upper left corner. (*Courtesy of P. I. Dee and C. W. Gilbert.*)

The high penetrating power of neutron radiations shows that neutrons do not interact noticeably with electrons. If such an interaction did exist, neutrons would ionize matter and thus lose energy. Actually the energy loss of neutrons is entirely due to direct collisions with nuclei. As can be easily shown from the laws of conservation of momentum, the energy loss per collision is greater in collisions with lighter nuclei. *Thus* *fast neutrons are very rapidly slowed down and scattered as they move through hydrogenous matter.* Eventually they acquire an average kinetic energy equal to that of the surrounding molecules at the given temperature (**thermal neutrons**). Obviously the thermal neutrons cannot be kept in any container. They have a tendency to diffuse through the walls of the container. But "dissolving" neutrons in a very large block of matter for storage is not feasible either. As we shall see later, they are captured by the surrounding nuclei (even those of hydrogen), giving rise to a heavier isotope of the surrounding element. So it seems that the only way to preserve free neutrons would be to shoot them into empty space in the form of a beam. But experimental evidence indicates that even this method would not preserve them indefinitely, since neutrons are radioactive and disintegrate with a half life of about $\frac{1}{2}$ hr by emission of beta rays, transforming themselves into protons! Thus neutrons appear to lead a permanent existence only in the interior of stable atomic nuclei, being only transient guests in the interatomic space.†

Positrons

While performing an experiment very similar to the one illustrated in Fig. 39.3 the Joliot-Curies made an interesting discovery. They subjected an aluminum target (located in the place of *Be* in Fig. 39.3) to bombardment by alpha particles of polonium. The reaction was similar to the one observed in the case of beryllium; the target was found to emit neutrons. But in addition to neutrons, they detected positively charged particles emanating from the target. Magnetic deviation revealed that these particles were not protons but rather positive corpuscles of the same specific charge e/m as electrons. This observation of positive electrons, or **positrons,** would have been quite startling had it been made several years earlier. But in 1934, when the Joliot-Curies made this observation, the existence of positive electrons was already well established owing to the work of Anderson, who observed such particles in a cloud chamber in 1932 in the course of his studies of cosmic radiation. Figure 39.5 shows the historic photograph which revealed the existence of positive electrons. His cloud chamber was placed in a magnetic field so as to enable him to distinguish between positive and negative particles by the direction of deviation and to estimate the momentum of the particle from the curvature of the path. In the case pictured in Fig. 39.5 the magnetic field is perpendicular to the plane of the paper. A faint track of an electron is seen in the picture to pass through a lead plate L placed across the Wilson chamber. The fact that the curvature of the track is greater below the lead plate shows that the particle came from above and was slowed down by passing the lead. The direction of the curvature indicates that it is a positive particle.

---

† The existence of neutrons in a stable nucleus is not strictly permanent according to contemporary nuclear theory. As has been mentioned in Chap. 38, a neutron and a proton are assumed to interact inside a nucleus by tossing a meson to and fro and thus changing the neutron into a proton by emitting a negative meson or absorbing a positive one and the proton into a neutron by losing a positive meson or absorbing a negative one. It is assumed that, whenever an unstable nucleus emits a beta particle, a neutron in the nucleus is transformed into a proton. Conversely, emission of a positron is imagined to be associated with conversion of a proton into a neutron.

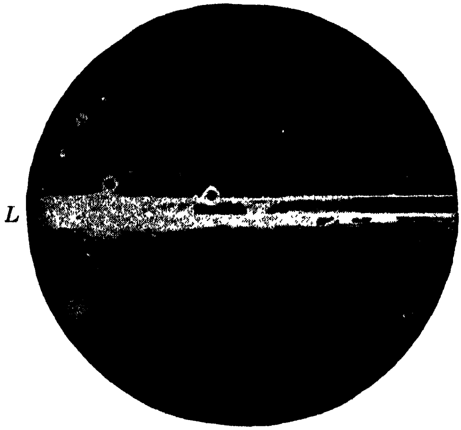


FIG. 39.5. Track of a positron in a Wilson chamber (L is a lead plate placed across the chamber). (*Courtesy of C. D. Anderson.*)

So far, we have no reason to suspect a revolutionary discovery. After all, this positive particle might be a proton or an alpha particle! The significance of this photograph was revealed by a quantitative consideration. Assuming the particle to be a proton, Anderson calculated from the curvature of its path in the known magnetic field the energy of the hypothetical proton. He found the value of 300,000 electron volts. On the other hand he knew from the empirical range-energy relationship that the range of protons of this energy in the atmosphere should be only 0.5 cm. But the length of the track photographed by Anderson was over 5 cm! He concluded that the observed particle was of a much smaller mass than a proton, presumably a positive counterpart of the negative electron. A particle of such a small mass would have a much greater velocity than a proton of equal energy. It would, therefore, act for a shorter time on the electrons of an atom which it passes and thus impart to them a smaller impulse and will be less likely to ionize the atom. It would thus lose energy by ionization less rapidly and have a longer range than a proton.

Physicists were not quite unprepared for the discovery of the positive electron. Anderson's photograph recalled in their memory a very strange theory of a young English theoretician Dirac, who predicted the existence of positively charged particles identical with the well-known negative electrons in mass and in the magnitude of their charge. Dirac's theory is very complicated, being based on mathematical considerations of quantum mechanics and the theory of relativity. Dirac gave a very imaginative interpretation of the implications of his so-called *relativistic wave equation.* In fact, he stretched the imagination of physicists to the utmost extent. We can give here only a very crude outline of some of the essential points of Dirac's theory with a warning to consider its pictorial content as nothing more than a provisional working hypothesis.

In discussing Bohr's theory, we assigned to a planetary electron certain "permissible" energy states. The electron possessed a certain kinetic energy and a potential energy which was defined so as to vanish at infinity and to have a negative value near the nucleus. The kinetic energy $\frac{1}{2}mv^2$ is naturally always positive, since m and v^2 are necessarily positive. The least energy a "free" electron could have is zero, namely, when it is at rest—at infinite distance from the nucleus. Dirac was, however, led by his wave equation to a startling conclusion. By interpreting both the negative and the positive root of a mathematical expression occurring in his theory, he was led to the assumption that there can exist free electrons of negative *kinetic* energy! It is difficult to see what physical meaning can be attached to a negative kinetic energy, and consequently we cannot exactly say that Dirac's theory can provide an explanation of physical phenomena if we mean, by the term "to explain," to describe in terms of more familiar concepts. But in spite of this seemingly irrational element of Dirac's theory it led to predictions which were confirmed by experiment, and it is presented here as an illustration of how far modern physics is willing to go in justifying radical assumptions as long as they lead to a successful scheme of prediction.

Dirac assumed the whole empty space to be infinitely densely populated with an infinite number of ordinary electrons (of negative charge), having all conceivable values of *negative* kinetic energy ranging *from* $-mc^2$ *to* $-\infty$ (where m is the mass of the electron at rest and c the velocity of light). According to Einstein's law the energy inherent in a free electron at rest is $E = mc^2$. According to Dirac's hypothesis *there are no free electrons whose energy ranges between* $+mc^2$ *and* $-mc^2$. But *all* values of electron energy are possible between $-\infty$ and $-mc^2$ and between $+mc^2$ and $+\infty$. The whole space is pictured as an "ocean" of imperceptible electrons containing an infinite number of electrons of negative kinetic energy. We have no way of observing them because space is supposedly uniformly filled with them. However, if we disrupt this continuity, we become aware of their existence. If we supply enough energy to one of them to "raise" it to a positive energy state, it becomes apparent as an ordinary negative electron whereas the "hole" in the "electron ocean" from which it has been raised becomes noticeable as a "lack of a negative electron of negative energy." *This "bubble" in Dirac's ocean of negative electrons of negative energy should behave, according to Dirac's hypothesis, like an ordinary electron of a positive charge* **(positron).** According to Dirac, a moving positron is a wandering "hole" in the ocean of electrons of negative energy. Such holes should not be long-lived owing to the tendency of ordinary electrons to "fall" into them, thus leading to the disappearance of both the positron and electron.

Dirac's prediction of the processes of *mutual annihilation of an electron and a positron* as well as his prediction of a *simultaneous creation of pairs* of electrons and positrons [*pair production* by knocking an electron out of Dirac's ocean, which leads to the observation of a negative electron and the positive "hole" (positron)] were confirmed by experiment. The experimental evidence will be discussed below.

Mesons

Bombardment of a target by alpha particles of unprecedentedly high speed (360 million electron volt alpha particles, accelerated in the Berkeley 184-in. cyclotron) has yielded new elementary particles, the so-called *mesons.* These artificially generated particles are the same as those which have been postulated by Yukawa to exist inside atomic nuclei and which have been observed in the cosmic radiation (see Chap. 38).

Where do the cosmic-ray mesons come from? We are certain that they do not come to us from interstellar space. Because of the very short half life of these unstable particles, which eventually decay emitting an electron and presumably neutrinos after an existence lasting on the average not much longer than a millionth of a second, they cannot cover large distances without changing their identity. They are assumed to be generated in the earth's atmosphere by impact of heavier primary cosmic-ray particles upon the nuclei of the atmospheric gases.

Figure 39.6 shows the record of a primary cosmic-ray meson (π) giving rise to a secondary cosmic-ray meson (μ) in a photographic emulsion. The emulsion is sensitive to heavy particles, such as protons and α particles, but not to electrons. Thus we know that the long faint tracks are not due to electrons but to heavier particles. The three short tracks shown in the lower quarter of the photograph are made by α particles. They are seen to be much denser than the meson tracks. The direction of motion of the mesons is inferred from the changes in the track density. As the meson slows down, the track becomes heavier. Thus, we see that the π meson enters at the top of the figure and the μ meson originates where the track of the π meson ends.

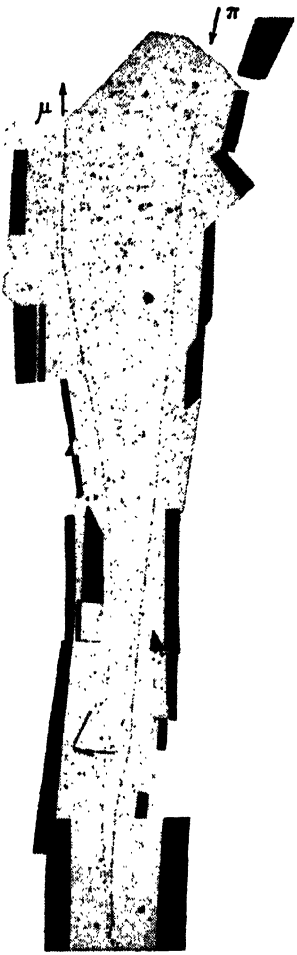


Fig. 39.6. Two meson tracks in a photographic emulsion. The meson π moves downward and gives rise to a second meson μ which moves upward. *(From Powell and Occhialini, "Nuclear Physics in Photographs," Oxford, Clarendon Press.)*

DISCOVERY OF NEW PROCESSES

Creation of Electron Pairs

The discovery of new elementary particles was accompanied by discoveries of new physical processes. Figure 39.7 shows a photograph obtained by sending gamma rays of thorium C'' through a cloud chamber which was placed in a magnetic field (whose field lines are perpendicular to the plane of the paper). The arrow marked γ indicates the direction of the gamma rays which pass through the cloud chamber. We see a pair of curved lines originating at a point inside the cloud chamber. Their opposite curvature indicates that they carry opposite electric charges. The density of the track produced makes them immediately recognizable as an electron and a positron (the polarity of each particle is indicated in the diagram). *Such "pairs" of electrons and positrons are observed only with gamma rays, whose photon energy exceeds* $h\nu = 1.02 \times 10^6$ *electron volts.* How can this observation be explained?

According to Dirac's "hole theory" this phenomenon is explained as follows: In order to make one of the electrons of "Dirac's ocean of electrons of negative energy" observable, we must raise it to a state of positive energy. We observe it then as a free negative electron, whereas the vacancy which it leaves in Dirac's ocean is observed as a positron. The highest energy which an electron can have in Dirac's ocean is $-mc^2$ (the lowest value being $-\infty$). The lowest value of positive energy which an electron can have is $+mc^2$ for an electron at rest. Hence, the minimum energy required to raise an electron from the highest negative energy state to the lowest positive energy state is $2mc^2$. Expressed in electron volts this energy amounts to 1.02×10^6 electron volts, which is precisely the minimum energy of gamma quanta which are known to produce "pairs." If $h\nu$ exceeds this value, the surplus energy is observed as the kinetic energy of the electron and positron. "Pair production" does not occur in a vacuum. It occurs only in the presence of matter, and the probability of the process increases with the nuclear charge of the material through which the γ radiation passes. *This process of creation of massive particles at the expense of the energy of a photon is often referred to as "materialization of energy." It offers a quantitative confirmation of Einstein's mass-energy law.*

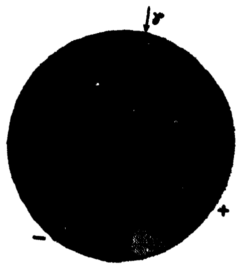


FIG. 39.7. Production of an electron-positron pair by γ rays. (*Recorded by Simons and Zuber. From White, "Modern College Physics," Van Nostrand.*)

Annihilation of Electron Pairs

The spectacle of creation of elementary material particles from radiation energy suggests the question whether the converse process of *conversion of mass into radiation energy* can be observed. We cannot expect to demonstrate this process as strikingly as that of "pair creation," since the gamma radiation which we can expect to be the result of the conversion of the mass of an electron into radiant energy does not produce ion tracks in the cloud chamber. We have, however, convincing indirect evidence of this process. Wherever there are positrons absorbed in matter, we observe the emission of gamma radiation of photon energy $h\nu = 0.51 \times 10^6$ electron volts. This can be interpreted as being due to conversion of the positron mass, which is equivalent to the energy of 0.51×10^6 electron volts, into energy of a gamma ray photon of 0.51×10^6 electron volts. According to the picture of Dirac's theory this process of "annihilation of electrons" can be visualized as the falling of an ordinary negative electron of positive energy into a "hole" in Dirac's ocean, *i.e.*, into a vacant state of negative energy which manifests itself as a positron. If the interacting particles are at rest, the energy emitted in this quantum jump is $2mc^2$. The law of conservation of momentum leads us to expect this energy to be emitted in the form of two gamma quanta of equal energy $h\nu = mc^2 = 0.51 \times 10^6$ volts proceeding in opposite directions, so that the combined momentum of these quanta is zero. The fact that one actually observes an "*annihilation radiation*" of a sharply defined energy per quantum of $h\nu = 0.51 \times 10^6$ electron volts indicates that positrons usually come to rest before annihilation. In this process the two opposite charges cancel each other. The complete disappearance of charge in the process of mutual annihilation of positive and negative electrons can be accepted as evidence of the exact equality of the charge of the positron to the charge of an electron.

Induced Radioactivity

In discussing the artificial production of positrons we described the experiment of the Joliot-Curies in which an aluminum foil was bombarded by alpha particles of polonium. That experiment resulted in the detection of two kinds of particles ejected as a consequence of nuclear transformations, neutrons and positrons. Did they discover a new process in which a nucleus emits two different particles simultaneously? They did not, but they discovered a new process of a different type, the so-called **induced radioactivity,** which is of great interest to the physicist as well as to the chemist and the biologist because of the revolutionary methods of research which can be based on it (see Appendix). Their discovery consisted in the observation that the emission of positrons continued when the polonium source was removed whereas the neutron emission stopped immediately. The foil had evidently been made radioactive (*induced radioactivity*). Can we conclude from this that aluminum was made radioactive? Chemical separation of the produced radioactive material shows that it is not aluminum. The radioactive substance is the element formed through interaction of an alpha particle with an aluminum nucleus. Since we know that a neutron is emitted in this process, we can write the following reaction:

$$ {}_{2}\mathrm{He}^{4} + {}_{13}\mathrm{Al}^{27} \rightarrow {}_{15}\mathrm{P}^{30} + {}_{0}n^{1} \qquad (39.9) $$

The phosphorus thus formed can be separated from aluminum by chemical methods. It is then seen that the positrons are emitted by phosphorus ${}_{15}\mathrm{P}^{30}$, which decays into silicon with a half life of about 2.5 min according to the reaction

$$ {}_{15}\mathrm{P}^{30} \rightarrow {}_{14}\mathrm{Si}^{30} + {}_{1}e^{0} \qquad (39.10) $$

(We adopt the symbol ${}_{+1}e^{0}$ for the positron and ${}_{-1}e^{0}$ for the negative electron. 0 is the mass number, *i.e.*, the nearest integer to the mass of the particle in amu.)

Since the discovery of induced radioactivity by the Joliot-Curies, hundreds of radioactive isotopes of various elements have been produced by artificial means. They can be formed, for instance, by bombarding nuclei with α particles, neutrons, or deuterons. The resulting radioactive nucleus emits in some cases electrons, in others positrons, depending on whether the unstable nucleus formed has an excess of neutrons or of protons, respectively.

In the case of a neutron excess in the nucleus a neutron is assumed to change into a proton under emission of a negative electron, whereas in the case of a proton excess a proton is assumed to change into a neutron under emission of a positron.

The "radioisotopes" are of great importance in so-called *radioactive tracer* studies. These radioactive elements have the same chemical properties as their nonradioactive isotopes. It is therefore possible to incorporate them into a well-known chemical compound (or food product) and then to follow the element, which has been "tagged" by the radioactivity of the admixed isotope, through a sequence of chemical reactions with the aid of a Geiger counter. In the case of food, the path of its tagged components can be followed through the body of the plant or of the animal. A more detailed account of radioactive tracer research will be given in the Appendix.

The following are additional examples of production of artificially radioactive isotopes:

bombarding particle	target	product	by-product	radioisotope	daughter nucleus	emission	
${}_1H^2$	$+\ {}_{15}P^{31}$	$\rightarrow {}_{15}P^{32}$	$+\ {}_1H^1$	$\therefore {}_{15}P^{32}$	$\rightarrow {}_{16}Si^{32}$	$+\ {}_{-1}e^0$	(39.11)
${}_1H^2$	$+\ {}_6C^{12}$	$\rightarrow {}_7N^{13}$	$+\ {}_0n^1$	$\therefore {}_7N^{13}$	$\rightarrow {}_6C^{13}$	$+\ {}_{+1}e^0$	(39.12)
${}_0n^1$	$+\ {}_{20}Ca^{42}$	$\rightarrow {}_{19}K^{42}$	$+\ {}_1H^1$	$\therefore {}_{19}K^{42}$	$\rightarrow {}_{20}Ca^{42}$	$+\ {}_{-1}e^0$	(39.13)
${}_0n^1$	$+\ {}_1H^1$	$\rightarrow {}_1H^2$	$+\ \gamma$†				(39.14)

Nuclear Transmutation by Neutrons

Reactions (39.13) and (39.14) are of particular interest to us. They indicate the possibility of using neutrons as projectiles for nuclear transmutation. The advantage of using neutrons for this purpose, instead of charged particles, such as protons, deuterons, of alpha particles, is immediately apparent. *Neutrons, having no charge, should be capable of entering nuclei without retardation by a coulomb repulsive force. We should expect them, therefore, to be the most efficient agents of nuclear transmutation.* This expectation is confirmed by experiments. **Practically all nuclei can be transmuted by neutron bombardment.** The reactions may involve an immediate transmutation of the element accompanied by emission of a heavy particle and the formation of a stable or unstable nucleus, such as in the following cases:

---

† This type of reaction is called *neutron capture.*

$$ {}_8O^{16} + {}_0n^1 \rightarrow {}_7N^{16} + {}_1H^1 \qquad (39.15)$$

$$ {}_5B^{10} + {}_0n^1 \rightarrow {}_3Li^7 + {}_2He^4 \qquad (39.16)$$

or they may consist of *capture of the neutron* and the formation of a radioactive isotope of the target nucleus, such as in this reaction:

$$\begin{array}{l} {}_{35}Br^{79} + {}_0n^1 \rightarrow {}_{35}Br^{80} + \gamma \\ \qquad\qquad\qquad\quad \downarrow \\ \qquad\qquad\qquad {}_{36}Kr^{80} + {}_{-1}e^0 \end{array} \qquad (39.17)$$

The bromine isotope formed by neutron capture is radioactive and decays into krypton under emission of negative beta particles.

The neutrons for such transmutation experiments must be produced by nuclear disintegrations in which they are ejected. One might, for instance, bombard a beryllium target with deuterons of one million electron volts and obtain neutrons according to the following reaction:

$$ {}_4Be^9 + {}_1H^2 \rightarrow {}_5B^{10} + {}_0n^1 \qquad (39.18)$$

or one may bombard "heavy ice" (*i.e.*, frozen "heavy water," a combination of deuterium and oxygen: D_2O)† with deuterons of several hundred thousand electron volts:‡

$$ {}_1H^2 + {}_1H^2 \rightarrow {}_2He^3 + {}_0n^1 \qquad (39.19)$$

The neutrons emitted in such processes are usually very fast. For instance, the energy of neutrons emitted in reaction (39.18) is about 3×10^6 electron volts.

In studying transmutations of nuclei by fast neutrons, Fermi and his coworkers made an important discovery. They observed that, when the source of fast neutrons and the target were surrounded by hydrogenous material, such as paraffin, the reaction yield (*i.e.*, the number of successful transmutations per neutron) was increased by a factor of about 10! They also found that interposing a piece of paraffin between the neutron source and the target greatly increased the effectiveness of the neutrons in causing nuclear transmutations. This observation was explained as follows: In passing through matter, neutrons collide with surrounding nuclei and lose part of their kinetic energy by accelerating the slower nuclei with which they collide elastically. According to the laws of elastic impact (derived from the laws of conservation of momentum and energy) the energy loss of a neutron per collision is the greater the smaller the mass of the nucleus with which it collides. Hence, such energy losses should be most conspicuous in hydrogen. It can be calculated that neutrons of 5×10^6 electron volts of energy will be slowed down to the average thermal velocity of the surrounding hydrogen molecules (corresponding to a molecular kinetic energy of $\sim \frac{1}{35}$ electron volt) after about 20 collisions. It is plausible that the slow neutrons thus formed should be more easily captured than fast neutrons. The slower a neutron the longer is the time during which it remains in the vicinity of a nucleus which it passes and the greater is, consequently, the impulse which the force of nuclear attraction (which we postulate to exist between neutrons and nuclei) exerts upon the passing neutron, and since, moreover, a slow neutron is more easily deviated from its path than a fast one, there will be a greater chance for a slow neutron to be pulled into a nucleus and to be captured by it. In the original experiment of Fermi and his collaborators, the surrounding of the source and target by hydrogenous material added the advantage of scattering back toward the target some of the neutrons which emanate from the source in all directions and which would otherwise not have reached the target.

---

† Deuterium, D, is "heavy hydrogen" ${}_1H^2$.

‡ This reaction can still be observed at much lower voltages.

Transmutation by slow neutrons is the most efficient process of nuclear transformation. Whereas, for instance, the "yield" in transmutations by 5×10^6 *electron volt deuterons in favorable cases is about one transmutation for* 10^5 *impinging deuterons, practically every slow neutron is eventually captured by a nucleus, thus giving rise to a nuclear transmutation.*

Synthesis of Transuranium Elements

The absence of repulsive forces between neutrons and nuclei suggested to Fermi the idea of attempting to synthesize a nucleus of greater atomic number than any nucleus existing in nature. The heaviest element known at the time, uranium, which consists of a mixture of two isotopes [$_{92}U^{238}$ (99.3 per cent) and $_{92}U^{235}$ (0.7 per cent)] was subjected to bombardment by neutrons. One would expect to observe the formation of a heavy uranium isotope $_{92}U^{239}$ (we neglect for the time being any possible transformation of $_{92}U^{235}$ which is present in very dilute form in the uranium mixture). Since all elements of atomic number above 84 are radioactive, it was reasonable to expect the anticipated uranium isotope to be also radioactive. Having a neutron surplus, this nucleus could be expected to decay by β^- emission into one of higher atomic number. A new radioactive isotope could be distinguished from the original radioactive nuclear species by its presumably different half life.

What Fermi and his associates found was more complicated than the simple anticipation presented above. They found in the uranium sample, after exposure to neutrons, four different products with the following half lives: 10 sec, 40 sec, 13 min, and 90 min. These half lives are quite different from that of uranium (4.51×10^9 years) and of its daughter elements. Subsequent research showed that "transuranium elements" (*i.e.*, elements having an atomic number exceeding that of uranium) are indeed formed under the conditions of Fermi's experiment, but the results are complicated by the presence of other radioactive products which display a variety of half lives.[†] The following sequence of transformations leading to transuranium elements was finally established:

$$ {}_0n^1 + {}_{92}\mathrm{U}^{238} \xrightarrow[\gamma\text{ emission}]{} {}_{92}\mathrm{U}^{239} \xrightarrow[\beta\text{ emission}]{\substack{\text{half life}\\ \text{21 min}}} \overset{\text{(neptunium)}}{{}_{93}\mathrm{Np}^{239}} \xrightarrow[\beta\text{ emission}]{\substack{\text{half life}\\ \text{2.4 days}}} \overset{\text{(plutonium)}}{{}_{94}\mathrm{Pu}^{239}} \xrightarrow[\alpha\text{ emission}]{\substack{\text{half life}\\ \text{30,000}\\ \text{years}}} {}_{92}\mathrm{U}^{235} \qquad (39.20) $$

The preceding chain of transformations shows how, with each nuclear electron emission, an element is formed whose atomic number exceeds by 1 that of its parent.

Subsequently transuranium elements of still higher atomic numbers have been synthesized. Americium ${}_{95}\mathrm{Am}^{241}$ is produced by bombarding ${}_{92}\mathrm{U}^{238}$ with α particles. The result is formation of plutonium according to the reaction

$$ {}_{92}\mathrm{U}^{238} + {}_2\mathrm{He}^4 \rightarrow {}_{94}\mathrm{Pu}^{241} + {}_0n^1 $$

This plutonium isotope decays by emission of beta rays into americium:

$$ {}_{94}\mathrm{Pu}^{241} \rightarrow {}_{95}\mathrm{Am}^{241} + {}_{-1}e^0 $$

The transuranium element curium can be synthesized by bombarding the lighter plutonium isotope with α particles:

$$ {}_{94}\mathrm{Pu}^{239} + {}_2\mathrm{He}^4 \rightarrow {}_{96}\mathrm{Cm}^{242} + {}_0n^1 $$

Both americium and curium decay by emission of α particles.

Nuclear Fission

The discovery of transuranium elements was, however, not the only result of the bombardment of uranium by neutrons. The most interesting discoveries resulted from efforts to separate from each other the elements of different half life, resulting from the irradiation of uranium by neutrons. This task was greatly complicated by the fact that the number of observed half lives grew from the four, originally observed by Fermi's group, to nine. These half lives could hardly belong to transuranium elements. It was very unlikely that the ${}_{92}\mathrm{U}^{239}$ nucleus, presumably formed by neutron capture, would give rise to such a long unbroken succession of beta emitters, since the increase in Z following each β^- emission would eventually lead to a nucleus of mass 239 and a charge of over 100. This would be at variance with the general tendency of the atomic mass of the elements to increase with the atomic number.

[†] The nature of these nontransuranic radioactive products will be clarified in the following section. The half lives listed above are not due to the transuranium elements.

The general confusion was greatly increased when Curie and Savitch announced that one of the elements (having a half life of $3\frac{1}{2}$ hr) in the sample of uranium which was bombarded by neutrons behaved similarly chemically to lanthanum ($Z = 57$). The conclusion seemed inevitable that it was actinium ($Z = 89$), whose chemical properties are very similar to those of lanthanum, rather than lanthanum itself.† To form ${}_{89}\mathrm{Ac}$, a ${}_{92}\mathrm{U}$ nucleus would have to lose only three positive elementary charges, but to give rise to ${}_{57}\mathrm{La}$, it would have to get rid of 35 positive charges!! There is no evidence for such a long chain of radioactive α or β^+ transformations. And yet Curie and Savitch were wrong in disregarding the results of the chemical analysis which indicated the presence of lanthanum! Had they had confidence in their chemical analysis and the courage to accept the striking result to which it seemed to point, they would have been credited with one of the most important discoveries of this century. They continued to experiment with the questionable product until they finally reached the decision that its chemical properties were somewhat different from those of lanthanum and ended by proposing the hypothesis that the "$3\frac{1}{2}$-hr product" was a transuranium element.

Their work was repeated by Hahn and Strassman. These investigators escaped the danger of permitting theoretical arguments to interfere with the recognition of facts. They prepared a radioactive precipitate from the dissolved sample of uranium which was previously exposed to neutrons. The method of precipitation established that it could have been either barium ($Z = 56$) or radium ($Z = 88$), which are chemically similar, being in the same column of the periodic table. Naturally, they suspected the product to be radium rather than barium for the same reason that Curie and Savitch suspected originally the $3\frac{1}{2}$-hr product to be actinium rather than lanthanum. But even though this suspicion seemed in some ways plausible, there were also reasons to doubt it. The most plausible assumption about the mode of transformation of ${}_{92}\mathrm{U}^{239}$ into ${}_{88}\mathrm{Ra}$ would be a succession of two α decays. But a careful search revealed no alpha-emitting products in the irradiated samples! They decided to establish beyond doubt, *by chemical means*, whether the radioactive component of the precipitate was radium or barium. Much to their surprise, they found that it was indeed barium.

---

† These two elements are in the same column of the periodic system.

But how do their findings jibe with those of Curie and Savitch? The radioactive barium of Hahn and Strassman emits electrons and is thus transformed into the element of the next higher atomic number, namely, the lanthanum which was observed by Curie and Savitch:

$$_{56}\mathrm{Ba} \rightarrow {}_{57}\mathrm{La} + {}_{-1}e^{0}$$

This discovery was revolutionary in more than one way. It demonstrated a new mode of nuclear disintegration, induced by neutron bombardment of uranium. Whereas previously only small fragments, such as protons or alpha particles, had been observed to be ejected from nuclei, the discovery of Hahn and Strassman seemed to reveal a splitting of the uranium nucleus into large fragments about half of the size of the original nucleus. This phenomenon has been called **nuclear fission.** A more detailed subsequent study revealed that the uranium nucleus, rendered unstable by neutron capture, is split into two nearly equal fragments, such as barium and krypton:

$$_{92}\mathrm{U} \rightarrow {}_{56}\mathrm{Ba} + {}_{36}\mathrm{Kr}$$

or antimony and columbium:

$$_{92}\mathrm{U} \rightarrow {}_{51}\mathrm{Sb} + {}_{41}\mathrm{Cb}$$

The confusing feature of this process is that there is more than one way in which the uranium nucleus appears to be able to split, so that there is a large variety of daughter nuclei. The fact that these light daughter nuclei are radioactive, emitting beta rays with different half lives, partly explains the great initial confusion which existed in this field of research. Additional confusion is apt to be caused by the fact that fission is not the only reaction which can be caused by capture of neutrons by a uranium nucleus. We have seen in equation (39.20) that an unstable uranium isotope may be formed which decays into plutonium by two successive β emissions.

Experiments with neutrons of different speeds showed that production of plutonium occurs with neutrons of about 25 electron volts whereas fission is produced by the very much slower "thermal neutrons"! How can this "velocity effect" be explained? It was found that the two reactions occur with different uranium isotopes. Production of plutonium occurs as a consequence of absorption of a neutron of intermediate speed (energy of 25 electron volts) by a nucleus of the abundant isotope $_{92}\mathrm{U}^{238}$ *whereas fission occurs as a consequence of absorption of a thermal neutron by the scarce isotope* $_{92}\mathrm{U}^{235}$.

From the difference between the mass of the $_{92}U^{235}$ nucleus and the sum of the masses of the daughter nuclei (including the masses of the neutrons emitted in the process of fission), it is possible to compute from Einstein's law the amount of energy which is liberated in the fission of a uranium nucleus. The result obtained is about 200 million electron volts! *This is the greatest amount of energy release encountered so far in any nuclear reaction.* (The energy of alpha particles emitted by naturally radioactive nuclei is of the order of magnitude of a few million electron volts.) This suggests that, if we ever tried to utilize nuclear energy, the fission of uranium would be the most favorable process to consider.

Perhaps the most striking and practically the most important feature of the process of nuclear fission is the fact that neutrons are emitted as the uranium nucleus falls apart into nearly equal fragments. It takes but one neutron to form an unstable uranium nucleus by capture of the neutron, but *when the unstable nucleus undergoes fission, three to four neutrons may be emitted.*

Chain Reactions Involving Neutrons

This prolific generation of neutrons opens the possibility of explosive transmutation of matter in bulk, accompanied by an enormous release of energy. Figure 39.8 shows the progress of such a **chain reaction.** The neutrons are shown as black circles, $_{92}U^{235}$ nuclei as white ones, and the heavy "fission fragments" as irregular areas. We assume, for simplicity, only two *extra* neutrons to be ejected in a fission process in the above diagram, *i.e.*, three neutrons in all. We see how, beginning with only one neutron, which hits a $_{92}U^{235}$ nucleus, we obtain in the "next generation" three neutrons which can cause fission of three $_{92}U^{235}$ nuclei, in the course of which nine neutrons are ejected, each of which may produce fission, so that a third generation of 27 neutrons is born, and so on, until within an imperceptibly short time interval a large proportion of uranium atoms has undergone fission. Since even the so-called *slow neutrons* collide with another nucleus within ~10^{-5}

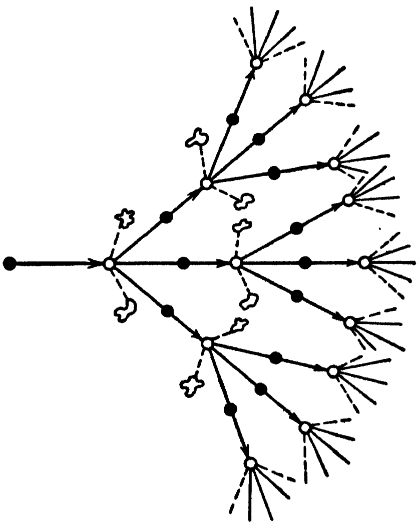


FIG. 39.8. Multiplication of neutrons in a chain reaction.

sec, an enormous number of nuclei undergo fission within less than a thousandth of a second. This process is accompanied by a release of energy per unit time which exceeds by far the effect of the most powerful chemical explosives. A simple computation shows that the energy released by fission of 1 gm of $_{92}U^{235}$ is about the same as that liberated in the explosion of 20 tons of TNT.

This suggests the idea of designing a powerful bomb using pure $_{92}U^{235}$ as explosive. But how could such a bomb be handled? It seems that it should blow up spontaneously, since one single neutron would be sufficient to start a chain reaction and since there are neutrons in the cosmic radiation which penetrates our atmosphere. But a mathematical analysis shows that an explosive chain reaction cannot take place unless the bomb has a certain critical size. The requirement of a critical size can be understood as follows. In describing the chain reaction above we assumed that every released neutron encounters and splits a uranium atom. This is not true in practice. The surface of the bomb constitutes a sort of window through which many neutrons escape. The smaller this "window" for a given volume the less will be the loss of neutrons. Since a sphere has the smallest surface for a given volume, the spherical shape will be the most advantageous. Since the surface of a sphere increases with the square of the radius whereas its volume grows with the third power, it can be seen that the surface-to-volume ratio diminishes as the sphere grows and hence a smaller proportion of neutrons can escape. The *critical mass* of a spherical sample of pure $_{92}U^{235}$ below which neutron losses due to escape through the surface are excessive, making the spread of a chain reaction impossible, is of the order of magnitude of 1 kg.

Thus a uranium bomb could be stored safely in two widely separated pieces of subcritical size. To explode the bomb the two pieces would be brought together very rapidly to form a chain-reacting critical mass. In order to release the energy in as short a time as possible, one utilizes fission by fast neutrons in the uranium bomb.

In addition to the escape through the surface, neutrons may fail to produce fission of $_{92}U^{235}$ by being captured in nonfissionable impurities. It is of particular importance to separate $_{92}U^{235}$ from the abundant isotope $_{92}U^{238}$, which is a nonfission captor of neutrons. The manufacture of the pure $_{92}U^{235}$ isotope represents one of the greatest difficulties of the uranium-bomb project. One of the successful methods of separating the two isotopes utilizes giant mass spectroscopes in which the path of moving uranium ions is bent in a magnetic field, the heavier of the two isotopes suffering a lesser deviation. The ions are caught in separate containers. Another successful method of isotope separation utilizes the fact that a gaseous compound of ${}_{92}U^{238}$ diffuses more slowly than the same compound of ${}_{92}U^{235}$.

${}_{92}U^{235}$ is not the only nuclear species which can undergo fission. It has been shown that other elements such as thorium and plutonium or the isotopes of uranium ${}_{92}U^{233}$ and ${}_{92}U^{238}$ are also "fissionable" (the latter by fast neutrons).

Chain Reactions Not Involving Neutrons

The neutron chain reactions are not the only type of possible nuclear chain reactions. Bethe has proposed an interesting chain of nuclear reactions as a hypothetical explanation of the source of the energy of the sun and possibly of stars in general. It is possible that physicists will be able to set off such chain reactions on earth sometime in the future.†

The essence of the Bethe chain is the conversion of hydrogen into helium. Four hydrogen atoms unite under ejection of two positrons and of three quanta of gamma radiation to form a helium nucleus. In this process 0.0275 amu disappears for each helium nucleus formed. This loss of mass is supposed to provide the necessary energy to maintain the interior of the sun (which loses energy at an enormous rate by radiation) at an estimated temperature of about 20 million degrees centigrade. At such extremely high temperatures the average thermal energy of the atoms is comparable to that acquired by electrons in an X-ray tube. Bethe assumes that, as the nuclei of the different elements collide at these temperatures, the following chain of reactions takes place:

$$\boxed{{}_{1}H^{1}} + {}_{6}C^{12} = {}_{7}N^{13} + \gamma \tag{1}$$

$${}_{7}N^{13} \rightarrow {}_{6}C^{13} + \enclose{circle}{{}_{+1}e^{0}} \tag{2}$$

$$\boxed{{}_{1}H^{1}} + {}_{6}C^{13} = {}_{7}N^{14} + \gamma \tag{3}$$

$$\boxed{{}_{1}H^{1}} + {}_{7}N^{14} = {}_{8}O^{15} + \gamma \tag{4}$$

$${}_{8}O^{15} \rightarrow {}_{7}N^{15} + \enclose{circle}{{}_{+1}e^{0}} \tag{5}$$

$$\boxed{{}_{1}H^{1}} + {}_{7}N^{15} = {}_{6}C^{12} + \enclose{circle}{\enclose{circle}{{}_{2}He^{4}}} \tag{6}$$

---

† One might, for instance, think of exploding a uranium bomb in a dense atmosphere of hydrogen and other elements which, at the high temperature created by the primary explosion, would undergo a nuclear chain reaction, such as the one described below.

The hydrogen nuclei which are consumed in this reaction are marked by square frames, whereas the helium nuclei and the positrons which are formed are marked by circles. The other nuclei which take part in this transmutation cycle play a "passive role." They are transmuted and resynthesized, so that their amount does not change, whereas the amount of helium increases at the expense of the hydrogen.

The Liquid-drop Model of the Nucleus

Is it possible to conceive of some kind of nuclear model which would enable us to explain on a common basis such widely different phenomena as spontaneous radioactive decay, disintegration by alpha particles, and nuclear fission? Bohr suggested the so-called *liquid-drop model* of the nucleus which has proved surprisingly successful in view of its extreme simplicity. Bohr imagines nuclei as drops of a sort of "nuclear solution" of protons dissolved in a neutron liquid. Since the protons repel each other, we must assume some force which counteracts their mutual repulsion in order to obtain a stable spherical drop of liquid. This is accomplished by assuming that neutrons attract protons as well as other neutrons. We then have conditions in our drop of nuclear liquid similar to those shown to cause surface tension in a liquid (Chap. 11). The mutual attraction between the neutrons and the attraction between the neutrons and protons cause the surface tension, whereas the mutual repulsion between the protons tends to diminish the surface tension.

A disturbance of the shape of the spherical drop will increase its surface and hence its surface energy. The drop will have a tendency to return to the spherical equilibrium shape of minimum potential energy. A transient deformation, elongating the drop into an ellipsoid, will cause the drop to perform oscillations in the course of which its shape is periodically deformed. The electrostatic forces of repulsion tend to elongate the ellipsoid, whereas the forces of surface tension tend to round it to a sphere. At great elongations the effect of the electrostatic repulsive forces becomes predominant. If the amplitude of elongation is high enough, the elongation of the drop begins to increase in an unstable fashion until the drop divides finally into two nearly equal drops. Figure 39.9 shows a fission of a liquid drop which vibrates with an excessive amplitude. In Fig. 39.9*b* a constriction has been formed owing to forces of surface tension, and in Fig. 39.9*c* the drop is shown to have split in two. Vibrations of the nuclear liquid drop are supposed to be set up by capture or by impact of a particle. In order to produce a large enough amplitude of vibration to cause fission, a certain minimum energy must be supplied to the nuclear drop.

This picture not only has been used to explain fission qualitatively but has even made possible surprisingly accurate quantitative predictions.

Bohr imagined the nucleons to be in lively motion, resembling molecular motion, inside the nuclear drop. The spontaneous emission of a nucleon he likened to the process of evaporation and the capture of a nucleon to the process of condensation. Just as a certain amount of heat is consumed when vapor molecules leave a liquid, a certain energy is required to eject a nucleon from a nucleus. And similarly, energy is released when a nucleon is captured, which is analogous to the heat of condensation in our liquid-vapor analogy. It is this "heat of condensation" or, rather, energy of capture which provides in the capture of slow neutrons the high energy necessary to set up the oscillations of the nuclear drop that lead to fission.

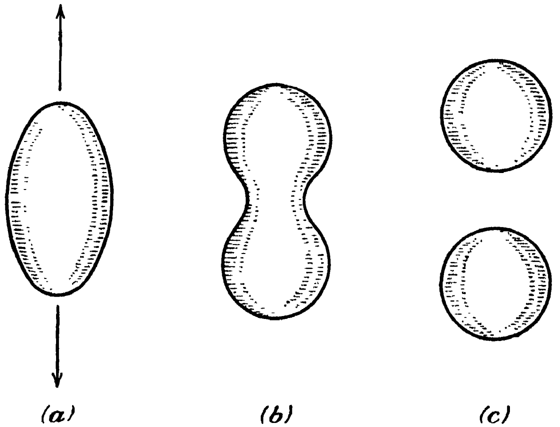


FIG. 39.9. Deformation of a vibrating drop which undergoes (a) elongation, (b) constriction, and (c) fission into two separate drops.

A certain average speed can be ascribed to the nucleons inside a nucleus just as to the molecules in a gas volume; we can, if we wish, speak of a nuclear "temperature" which can be defined in terms of the rms velocity of the nucleons. The speed of an individual nucleon will vary in time somewhat like the speed of a molecule. There is a certain chance that it may suffer several successive collisions which accelerate it and thus may accumulate enough kinetic energy to escape from the nucleus. We shall then have a case of spontaneous radioactive decay.

In the case of artificial transmutation we can imagine the bombarding particle to plunge into the target nucleus and to raise the interior "nuclear temperature" by collision with other nucleons. The general increase of "nuclear temperature" does not have to lead to an immediate "evaporation" of nucleons, but the increased rate of their agitation and their increased average energy make more probable an event in which one of the nucleons acquires by a "lucky" sequence of collisions an energy which is high enough to enable it to escape.

At this point the student has a right to launch a protest. We have shown convincingly that Newtonian mechanics does not hold for elementary particles moving in confines of atomic dimensions of the linear order of magnitude of 10^{-8} cm, and yet we are transferring the picture of the kinetic molecular theory, which is based on Newtonian mechanics, to a space of the linear order of magnitude of 10^{-12} cm! This objection is perfectly justified, and that is why nobody takes this model literally. It is merely a crude analogy which is based on a similarity, not of observable structural, but rather of energetic conditions in the compared processes. The fact that we have no better theory to present illustrates well the present lack of an adequate theory of nuclear structure.

APPENDIX

1. Ion Accelerators

A variety of machines have been developed for the purpose of accelerating ions for artificial nuclear transmutations.

Transformer Generators. Cockroft and Walton used high-voltage transformers in conjunction with condensers and rectifiers in a "voltage-multiplying" circuit. Insulation difficulties make such circuits impractical for voltages of the order of millions of volts.

Electrostatic Generators. Voltages of several million volts have been successfully generated by generators based on mechanical transport of charges. In the van de Graff machine, a silk conveyor belt moving at about 60 mi/hr is sprayed with electrons which are transported by the belt into the interior of a big metal sphere, where they are transferred to the surface of the sphere. The advantage of such generators is due to the fact that they provide a very steady voltage which can be measured precisely, whereas the voltage derived from rectified transformer circuits is fluctuating periodically. The production of a constant, precisely determined voltage is of importance for nuclear research.

The Cyclotron. The voltage applied to the vacuum tube of an ion accelerator cannot be increased indefinitely. At potential differences of several million volts it becomes very difficult to avoid sparks and "corona" discharges through air as well as breakdown of the glass insulation of the accelerator tube. Thus it was realized that one could hope to obtain particles with energies exceeding several million electron volts only by inventing ways of making the same particle "fall" repeatedly through a relatively low potential difference. This was accomplished in an ingenious fashion by Lawrence in his cyclotron as follows:

Figure 39.10 shows a schematic diagram of the accelerating circuit of the cyclotron in top view. Figure 39.11 shows a side view of the same parts. D_1 and D_2 are two electrodes in the shape of the letter *D*, the so-called *Dees*. Their shape and configuration can be visualized by imagining a closed pillbox cut in two and the two halves pulled slightly apart. The Dees are placed in a magnetic field (of the order of magnitude of 20,000 oersteds) which is perpendicular to the plane of the paper in Fig. 39.10.

It is generated by a huge electromagnet (Fig. 39.11) which weighs hundreds of tons. C_1 and C_2 are the coils of the electromagnet which are wound around its pole pieces N and S. The poles of the latest giant cyclotron at Chicago have a diameter of 200 in.

When an ion of charge e emanates from an ion source at I and moves in the plane of the paper (Fig. 39.10), it is forced to move in a circular orbit owing to the Lorentz force $F = Hev$ (Chap. 23) which is exerted upon it at right angles to its velocity vector and acts as a centripetal force. If we equate this expression for F to the product of mass times the centripetal acceleration, we get a very interesting result:

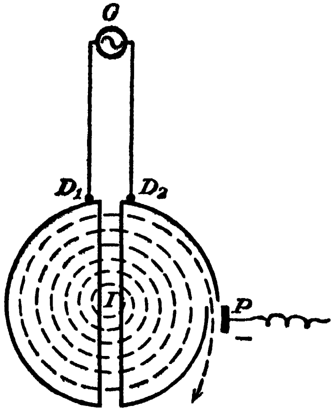


FIG. 39.10. Path of ions in a cyclotron (dashed spiral). D_1 and D_2, the *Dees;* G, the source of high-frequency alternating accelerating voltage; P, the deflecting plate which pulls the ions out of the Dee; I, the source of ions.

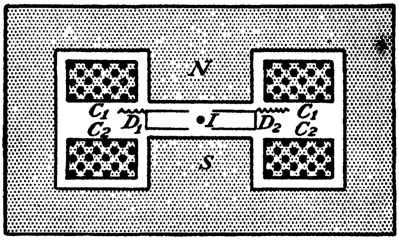


FIG. 39.11. Side view of the cyclotron. N and S, the poles of the magnet; C_1 and C_2, coil wound about the pole pieces; D_1 and D_2, the Dees; I, ion source.

$$Hev = m\omega^2 R$$

where $\omega = 2\pi f$, f being the number of full circles the ion completes each second and R the radius of its orbit. Substituting on the left side $v = \omega R$ we get

$$He\omega R = m\omega^2 R$$

hence,

$$\omega = H\frac{e}{m} = 2\pi f$$

or

$$f = \frac{H}{2\pi}\frac{e}{m} \tag{39.21a}$$

The time the ions require to complete a full circle is $T = 1/f$.

Hence,

$$T = \frac{2\pi}{H}\frac{m}{e} \tag{39.21b}$$

We see from this that the *time of revolution is independent of the radius of the orbit* for an ion of a given specific charge e/m moving in a constant magnetic field H.

This relationship makes the following trick possible: We apply to the Dees a high-frequency alternating voltage derived from the generator G. Imagine a particle moving on a circle about the point I as center. If at a moment when it moves at the lowest point of its orbit through the gap between the Dees from D_2 to D_1 the alternating voltage is in such a phase that D_2 is positive with respect to D_1, the (positive) ion is accelerated, and if the voltage amplitude of the potential difference across the Dees is 200,000 volts, the particle will gain, at best, 200,000 electron volts of energy. Now as the particle completes a semicircle and is ready to reenter the gap between the Dees, moving this time from D_1 to D_2, it will be decelerated if the potential difference between D_1 and D_2 has not changed. But we are at liberty to apply an alternating voltage to the Dees of such a frequency that, by the time the ion reaches the gap for the second time, the polarity of the Dees is reversed and the ion is accelerated again, gaining an additional 200,000 electron volts of energy. After completion of the next semicircle it will reach the gap again, and since the polarity of the Dees has been reversed by this time, it will be accelerated once more, etc. Thus we see that by adjusting the frequency of the a-c generator to equal the frequency of the ion in its orbit as given in equation (39.21*a*) we can accelerate the same ion repeatedly by a relatively low voltage and make it gain tremendous energies. After 50 revolutions an ion will gain 20 million electron volts of energy in the above example. Latest designs of cyclotrons (synchrocyclotrons)—with modifications necessitated by the increase in mass of the ions at very high speeds, which is predicted by the theory of relativity—yield particles of hundreds of millions of electron volts of kinetic energy. Giant cyclotrons are not only powerful research tools but also (small-scale) manufacturing plants for a variety of radioactive nuclei which can be produced by exposing various elements to the beam of fast ions emerging from the cyclotron. Such artificially produced "radioisotopes" find extensive use in medical research, treatment, and diagnosis and in biological and chemical investigations.

2. Neutron Sources

The Polonium-Beryllium Source. In cases where a portable source of neutrons is required, a mixture of polonium with beryllium powder can be used. The beryllium nuclei are disintegrated by the alpha particles of polonium according to the following reaction:

$$_2He^4 + {}_4Be^9 \rightarrow {}_6C^{12} + {}_0n^1$$

Deuteron-Deuterium and Deuteron-Beryllium Sources. If an ion accelerator of relatively low voltage (about 100,000 volts) is available, the following process can be used to obtain fairly high yields of about 3×10^6 electron volt neutrons according to the following "exothermal" reaction:

$$_1H^2 + {}_1H^2 \rightarrow {}_2He^3 + {}_0n^1$$

The target consists of heavy ice (frozen D_2O) which is bombarded by deuterium ions ("deuterons").

The most efficient source of neutrons, if deuterons of energies above 10^6 electron volts are available, consists of a beryllium target bombarded by deuterons. With 5×10^6 electron volt deuterons, neutrons of up to 9×10^6 electron volts of energy are obtained according to the reaction

$$_1H^2 + {}_4Be^9 \rightarrow {}_5B^{10} + {}_0n^1$$

The Uranium Pile. The fission of uranium ($_{92}U^{235}$) is utilized for the production of enormous quantities of neutrons in the so-called *uranium pile*. These neutrons can be utilized for large-scale transmutation of elements. *A pile is the most efficient method of manufacturing radioactive isotopes of various elements on a large scale.* The principle of the uranium pile is based on the following considerations:

We have seen that a sufficiently large piece of pure $_{92}U^{235}$ would explode in an uncontrollable chain reaction. What steps could be taken in order to regulate the rate of fission of this isotope so that the nuclear disintegration would proceed at a constant, moderate rate, yielding neutrons as a by-product? We have seen that several neutrons are released for each U^{235} nucleus which undergoes fission. For a self-sustained chain reaction it would be sufficient if only one of the fission neutrons released by each splitting uranium nucleus were captured by another U^{235} nucleus to cause its fission. The rest of the neutrons would then be available for other purposes, such as the manufacture of plutonium from U^{238} by neutron capture [see reaction (39.20)], or for manufacture of other nuclear species by capture of the extra neutrons. Some of the extra neutrons would unavoidably escape through the surface of the chain-reacting mass. Losses due to escape can be minimized by (*a*) using a large chain reactor which has a small surface-to-volume ratio, (*b*) by surrounding the chain reactor with a substance which reflects an appreciable proportion of neutrons back into the pile, and (*c*) by utilizing the escaping neutrons by surrounding the chain reactor with a "blanket" of a material which one desires to transmute by neutron capture.

Another source of neutron loss is nonfission capture in various impurities contained in the installation. Extreme care in purifying most of the materials used in a pile is one of the difficulties encountered in the construction of chain reactors.

In piles as opposed to bombs one uses slow (thermal) neutrons to induce the fission of U^{235}. To slow the neutrons down by collision, hydrogen would be the most effective element, but it is undesirable for another reason. It has a tendency to capture neutrons to form deuterons, thus reducing the number of neutrons available for the intended reaction. Deuterium would be good, but it is too difficult to obtain in sufficient quantities. Carbon in form of highly purified graphite proved the best practical material for slowing down the fast neutrons released in fission. A pile of graphite bricks is arranged as shown in Fig. 39.12. The bricks are provided with channels into which sealed aluminum containers containing the ordinary natural uranium mixture of U^{238} with U^{235} are inserted (U in Fig. 39.12). Fast neutrons emitted from atoms in one container will travel to uranium samples in other containers through the intervening bricks of graphite. Most of them will be retarded to "thermal" velocities by the time they reach uranium atoms in neighboring containers. Thus both uranium isotopes will be exposed to fast neutrons originating in the same container as well as to slow neutrons arriving from other containers.

The capture of slow neutrons by U^{235} is the vital reaction which leads to fission and thus maintains the chain reaction in the pile.† The capture of neutrons of intermediate speed by U^{238} nuclei is an unavoidable but useful reaction. It leads to generation of plutonium from the U^{239} isotope so formed [see reaction (39.20)]. Plutonium itself is valuable because it is fissionable. Its usefulness is enhanced by the fact that it can be separated from the mixture by *chemical* methods which are much less laborious than the *physical* isolation of the fissionable U^{235}. Pure plutonium can be used to make fission bombs or as a material for small piles. Plutonium piles can be made much smaller than the uranium pile described above because fission of Pu^{239} by fast neutrons can be utilized. This makes it unnecessary to introduce graphite for slowing down the neutrons, which reduces the size of the chain reactor.

Enough extra neutrons are available in a pile to allow one to introduce into some of the channels in the graphite bricks (see Fig. 39.12) various elements which are to be transmuted as a result of neutron capture. This method makes it possible to manufacture new radioactive nuclear species on a large scale. The radioisotopes thus produced find many uses in physical, chemical, biological, and medical research.

---

† Some U^{238} nuclei capture a fast neutron and undergo fission, but this reaction is rare and unimportant.

A very important feature of the pile is the mechanism which controls the rate of the chain reaction. One makes use of the fact that cadmium is the most efficient captor of slow neutrons. Cadmium rods whose depth of insertion can be controlled are inserted into a number of channels which traverse the pile (C in Fig. 39.12). When the cadmium rods are inserted all the way, the "death rate" of neutrons due to absorption in cadmium is too high for a chain reaction to proceed. The chain reaction can be started by pulling out the cadmium rods, and the rate of the chain reaction can be controlled by varying the depth of immersion of the cadmium absorbers.†

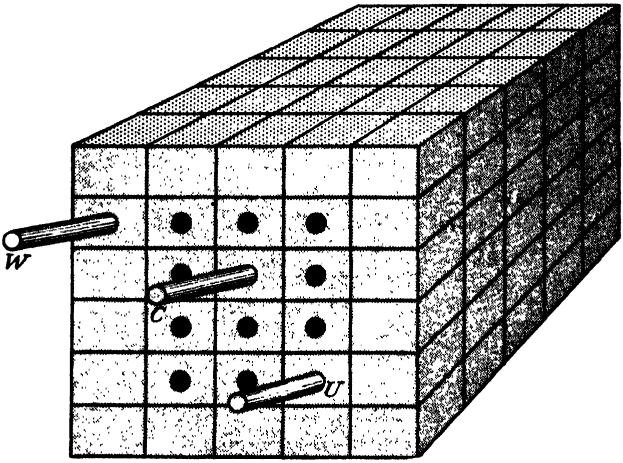


Fig. 39.12. Scheme of a uranium pile. Some graphite blocks contain channels into which airtight aluminum containers holding a natural mixture of uranium isotopes are inserted (U). Other channels contain cadmium rods (C), Geiger counters, containers filled with materials which are to be exposed to neutrons, and pipes (W) through which the cooling water is circulated.

After the chain reaction has been going for several hours, the aluminum cylinders can be removed (by remote control! since the intense radiation emanating from the pile and from irradiated samples is lethal to human beings) and the plutonium can be separated chemically from the uranium residue.

The nuclear chain reaction is accompanied by a tremendous release of energy. Eventually practical use will be made of the energy generated by chain reacting piles. At the moment this energy is wasted by heating large amounts of water which are passed through the pile (see pipe W which is one of the cooling pipes) in order to keep its operating temperature reasonably low.

---

† The fact that regulation of the chain reaction is possible at all depends on the occurrence of *delayed* neutron emission (*i.e.*, some neutrons may be emitted as late as a minute after the fission of the nucleus) which prolongs the time of the "response," making it possible to push in the cadmium rods before the reaction gets out of control.

3. Atomic Energy (Nuclear Energy)

The kinetic energy of the fission fragments (including the neutrons) and the energy released in the disintegration of the radioactive fission fragments are the sources of the heat developed by the pile. Due to the enormous amount of energy released in the fission of uranium (200 million electron volts per atom) a uranium pile is a very rich source of energy. (The fission of 1 gm of uranium liberates 20 billion calories of energy.) The utilization of power released by chain-reacting piles opens interesting technological possibilities. It may become economically worth while to generate electric power by nuclear fission in regions where water power is not readily available. Small chain-reacting piles (using "enriched" material, *i.e.*, one in which the concentration of the fissionable component is high) may propel ships and airplanes in the future. This development is of particular interest for nonstop flights (or for interplanetary rockets), since the mass of the nuclear fuel necessary to provide the required energy for the flight is negligible as compared with the amount of chemical fuel required. It is no exaggeration to say that the mastering of nuclear chain reactions may become as revolutionary an event in the development of our civilization as was the harnessing of fire by our ancestors. This development must be expected to have, directly or indirectly, a profound influence on the development of technology as well as on the progress of all natural sciences, provided scientists are given the opportunity to concentrate their efforts in that direction.

These prospects appear as a ray of hope in view of the energy famine which threatens the human race. Our energy resources, consisting of deposits of coal and oil, which have been stored over periods of millions of years, are wasted lavishly in relatively few years in times of war. At the present rate of consumption they cannot be expected to last more than a few thousand years. But if the fission of uranium will supply us with energy for the rest of the existence of our species, we do not have to worry. Unfortunately, however, our known deposits of uranium represent about the same store of energy as the store of chemical fuels, so that the crisis is only postponed by a few thousand years by utilization of uranium fission but is not averted indefinitely.

What will happen to our technological era in the absence of these fuel resources? Shall we have to revert to a primitive existence? Not necessarily. Several thousand years give us enough time to learn to harness the energy of solar radiation, which is quite considerable. We receive energy from the sun at the rate of 2 cal/(min)(cm^2). The power of the solar radiation impinging upon a square mile is about 2 million kilowatts, which is in the order of magnitude of the output of a large hydroelectric installation.

But the utilization of solar energy will not really be a revolution in our energy economy. Practically all our present sources of energy are "canned" solar energy. It is the energy of the sun which evaporates water from the lakes and oceans and raises the heated, humid air to higher strata of the atmosphere from which raindrops fall to the ground and eventually find their way to large rivers and waterfalls from which we derive hydroelectric power. It is the energy of the sun again which is absorbed by the chlorophyll of the plants, enabling them to synthesize substances essential to their own lives and to the lives of the herbivorous animals. Carnivores which eat the herbivores derive their energy thus indirectly from the sun. And finally our deposits of coal and oil are nothing but residues of ancient life derived from plants and animals that were buried millions of years ago deep underground and were transformed under the high-pressure and -temperature conditions at great depths into solid and liquid fuel.

What did we just show? Did we show that nuclear reactions will not play a decisive role in the final stages of human civilization? On the contrary! It was shown previously in this chapter how the high temperature of the sun is assumed to be maintained by the "Bethe cycle," which is a nuclear chain reaction in which hydrogen is converted into helium. It is a remarkable fact that this reaction has been proceeding at a substantially constant rate for hundreds of millions of years and at a rate which maintains on the surface of the earth a temperature appropriate for maintenance and evolution of life, which we have been able to follow from its early beginnings through the study of fossils. Thus we are led to the unexpected conclusion that life on earth owes its existence in part to nuclear reactions occurring on a lifeless star.

4. Applications of Radioactivity to Biology and Medicine

The introduction of the radioactive tracer technique into biological research is undoubtedly one of the most important methodological developments in contemporary biology. The migration and distribution of hormones and food in the body of plants and animals can now be studied, and facts can be ascertained about metabolism which could not be established by any other method. The food or drug whose path in the body is to be studied is prepared from radioactive materials. It is possible to manufacture nowadays radioactive isotopes of practically all elements, many of which have a long enough half life to be used in prolonged biological studies. Complicated organic compounds containing radioactive atoms can be manufactured in a variety of ways. For instance, radioactive starch can be obtained from plants which are grown in an atmosphere of $C^{14}O_2$ (C^{14} is a radioactive carbon isotope of a half life of 5,100 years). Plans are being made for a "radioactive farm" where radioactive inorganic materials will be used as plant food in an enterprise of growing radioactive plants. Feeding such plants to animals, such as cows, will produce radioactive livestock which could be used as radioactive food in tracer studies of nutrition and as a source of radioactive hormones.

There are two main methods used in the analysis of such experiments. The first consists in localizing the radioactive material in the living body by a small Geiger counter which can detect beta and gamma rays. The second consists of killing the animal, excising the tissue which is to be examined, and placing a section on a photographic plate. The radioactive material accumulated in the tissue will then blacken the photographic plate in places where it is concentrated (*radioautography*). Figure 39.13 shows, for instance, a radioautograph of a section of a tomato plant which has absorbed radioactive zinc from the soil. The zinc is seen to be concentrated in the seeds. The same technique can be used with animal tissues.

The following additional examples will further illustrate how radioactive tracers can be used in biological research.

a. When radioactive sodium Na^{24} is swallowed in the form of NaCl, a Geiger counter held near a fingertip begins to record beta and gamma rays about 2 min after the liquid reaches the stomach. By this method it can be shown that the salt rapidly enters the blood stream, by which it is transported to all parts of the body.

b. Feeding radioactive phosphorus to young growing rats shows that phosphorus is absorbed mostly in their bones. But interestingly enough one finds that even the bones of adult rats which do not grow become radioactive. This indicates that an exchange of atoms takes place between the ingested food and a tissue which does not grow, such as the bones. This information could not have been ascertained by any other method except the use of nonradioactive isotopic tracers.

c. It was found that in hyperthyroidism the thyroid gland absorbs iodine extremely rapidly and also has a tendency to lose it very rapidly. This symptom may find clinical application in the diagnosis of toxic goiter.

The location of the absorbed iodine in the thyroid tissue has been investigated by feeding patients radioactive iodine before a thyroidectomy. A section of the excised gland was then placed on a photographic plate, and the distribution of iodine was revealed in the radioautograph.

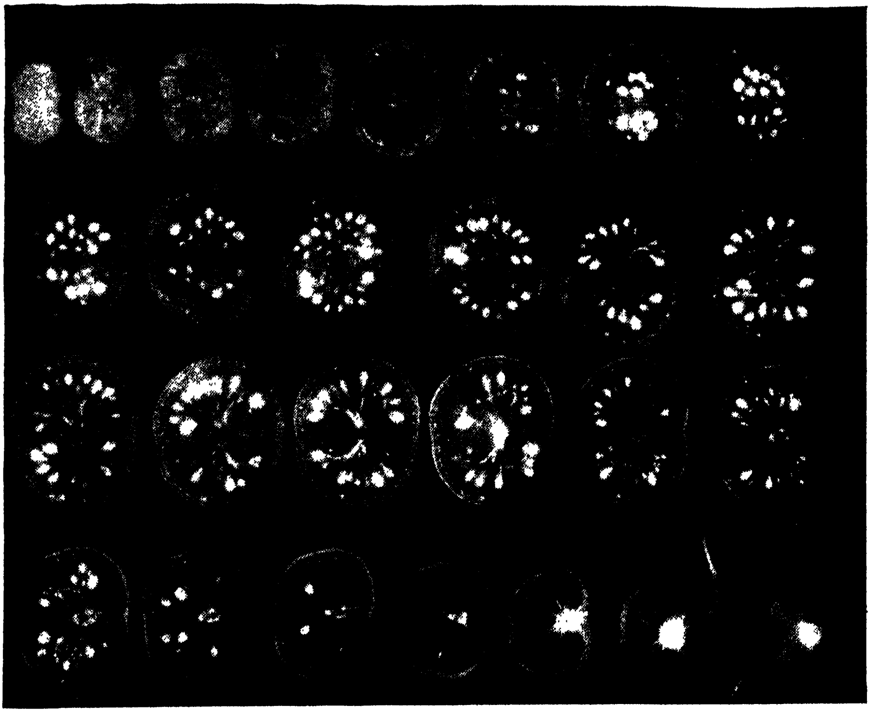


FIG. 39.13. Radioautographs of a tomato plant which has been fed radioactive zinc chloride. The white areas are regions where the photographic plate (the negative) has been blackened by the β rays of the radioactive zinc. (*Courtesy of Perry Stout.*)

d. The selective concentration of iodine in the thyroid gland can be used in the treatment of hyperthyroidism as an alternative to surgery or X-ray treatment. An appropriate dose of radioactive iodine is administered to the patient in order to kill a portion of the hyperactive thyroid tissue by radiation. The treatment stops automatically when nearly all the radioactive iodine atoms have decayed.

e. Administration of radioactive strontium by mouth to patients suffering from bone sarcoma showed by means of radioautographs that strontium is selectively absorbed in such tumors. This observation offers hope of more effective treatment of such malignancies, since the radioactive strontium would concentrate at the site of the proliferation and damage it by radiation without affecting too much of the healthy tissue.

f. A much-used treatment of cancer consists in implanting "radon seeds" (hollow, thin-walled, hermetically sealed metal needles filled with radon gas) into the diseased tissue. This does not provide a uniform irradiation of the tissue by γ rays. The present trend of research in this field is to search for substances which would be selectively absorbed by cancerous tissues. Such selectively absorbed drugs could be prepared from radioactive materials so that they would irradiate uniformly and kill the entire tumor in which they have been selectively concentrated.

g. Selective absorption of drugs in tumors is also of diagnostic value. Recently some dyes have been found which are absorbed selectively in brain tumors. These dyes have been prepared from radioactive materials so that the site of the brain tumor can be located with the aid of a Geiger counter. This method is important because localization of brain tumors by X rays is very difficult and sometimes impossible and the use of the electroencephalograph for this purpose is complicated.

h. Neutrons are very effective in killing tissues, and their suitability for killing cancerous growths is being studied. Slow neutrons are about five times as effective in killing rats as X rays. A rat can be killed within 2 min by exposure to a neutron beam produced with the aid of a cyclotron. Neutrons are also about five times as effective as X rays in killing cancerous growths in vitro. Their use in cancer therapy will depend on results of investigations of their effect on tumors in vivo.

QUESTIONS AND PROBLEMS†

1. Is the transmutation of nuclei a chemical or a physical process?

2. Does the emission of certain particles in a radioactive process necessarily prove that those particles are present in the nucleus?

3. Determine as nearly as you can from the data given in this text the answers to the following questions. How much energy is liberated when the following masses are completely converted into energy: mass of (*a*) an electron, (*b*) a proton, (*c*) a U^{235} atom, (*d*) 1 gm of U^{235}, (*e*) 1 gm of hydrogen.

4. How much energy would be liberated by formation of a nucleus of oxygen O^{16} from neutrons and protons?

5. Assuming that a noticeable nuclear transformation of deuterons takes place when 10,000 electron volt deuterons collide with deuterons at rest, estimate roughly at what temperature you would expect to observe nuclear disintegration in a hot deuterium gas. (Assume, for simplicity, that all atoms move with the rms velocity at a given temperature.)

6. Which will travel farther in air, an alpha particle or an electron of equal energy? Which one will produce more ion pairs?

7. Must a proton have a definite minimum energy in order to cause transmutation of a nucleus?

8. Complete the following nuclear reactions:

a. ${}_1H^1 + {}_3Li^7 \rightarrow {}_2He^4 + \cdots.$

b. ${}_1H^2 + {}_6C^{13} \rightarrow {}_6C^{14} + \cdots.$

c. ${}_1H^2 + {}_4Be^9 \rightarrow {}_5B^{10} + \cdots.$

d. ${}_0n^1 + {}_{35}Br^{79} \rightarrow \cdots \rightarrow \cdots + {}_{-1}e^0.$

e. ${}_{15}P^{30} \rightarrow {}_{14}Si^{30} + \cdots.$

9. Design an apparatus for detection of (*a*) fast, (*b*) slow neutrons (1) by a counter, (2) by producing a luminous effect on a screen.

---

† See Table I, p. 867.

10. A gamma ray of 1.4×10^6 electron volt energy is absorbed, giving rise to an electron and a positron of equal speed. How fast do the particles move?

11. Deuterium can be split by gamma rays into a neutron and a proton ("nuclear photoeffect"). If the mass (in amu) of the deuterium *atom* is $m_D = 2.0147$, what is the minimum energy the gamma ray must have to split the *nucleus*?

12. Compute (approximately) the velocity imparted to the proton in the photodisintegration of the deuteron (see Prob. 11) by a γ ray quantum of 3×10^6 electron volts of energy.

13. Compute the energy (in millions of electron volts) of the α particles in the following nuclear reaction:

$$_3\text{Li}^6 + {}_1\text{H}^2 \rightarrow {}_2\text{He}^4 + {}_2\text{He}^4$$

Assume the following *atomic* masses (in amu):

$$\begin{aligned} _3\text{Li}^6 &= 6.0169 \\ _1\text{H}^2 &= 2.0147 \\ _2\text{He}^4 &= 4.0038 \end{aligned}$$

Assume the energy of the deuteron to be 10^6 electron volts.

14. Using equation (39.7) determine (approximately) the velocity imparted by 10^6 electron volt neutrons (*a*) to protons in a head-on collision, (*b*) to nitrogen nuclei in a head-on collision.

15. Determine the amount of heat in calories produced for each gram of $_2\text{He}^4$ formed in the Bethe cycle (see section on Chain Reactions Not Involving Neutrons).

CHAPTER 40

THE THEORY OF RELATIVITY

SUMMARY†

Experiments designed to reveal an interaction between the ether and material bodies moving through it led to contradictory results. Some observations seemed to show clearly that the ether penetrates solid bodies and is not dragged along with a body which moves through it. Other experiments appeared to demonstrate equally clearly that a body moving through space drags the ether along with a speed smaller than its own speed relative to the ether. The famous experiment of Michelson and Morley finally showed with great precision that there is no "ether wind" near the surface of the earth, which moves about the sun with an orbital speed of about 18.5 mi/sec "relative to the ether." These contradictory results led to the conclusion that the problem of measuring the velocity of a body relative to the ether has no physical meaning. The problem of having to explain the mutually inconsistent results was solved by assuming that there is no such thing as a "luminiferous ether." The impossibility of detecting the motion of the earth relative to the ether (result of the Michelson-Morley experiment) was thus interpreted as the demonstration of the impossibility of detecting by optical experiments the absolute uniform linear motion of a body through space. Electrical experiments aimed to detect the motion of the earth through space also failed.

Einstein accepted the results of these experiments, along with the long-known fact that all mechanical processes proceed in identical fashion in systems which move relative to each other with a constant velocity, as indicating that *all physical processes* proceed according to the same laws in systems which move with a constant velocity relative to each other. Because of this no physical experiments of any kind performed in a closed system could reveal its uniform relative motion with respect to another system. He adopted as a basic postulate the following statement: **All physical processes proceed according to the same laws in all systems of reference which move relative to each other with a constant velocity (principle of relativity).** In particular the fact that the velocity of light in a vacuum is $c = 3 \times 10^{10}$ cm/sec was considered as a law of physics which holds in all such systems. (If the velocity of light were to vary from system to system, its variation would make it possible to determine the velocity of the closed system without reference to outside bodies. This, as we know from experience, is impossible.)

---

† In a less intensive course the study of this chapter may be replaced by a review of this summary.

Certain astronomical observations led to the conclusion that the velocity of light in a vacuum is the same regardless of the relative velocity between the light source and observer. For example, the velocity of light issuing from one of the two revolving stars of a double star is 3×10^{10} cm/sec during the period when the star recedes from the earth as well as during the time when it moves toward us. Einstein adopted this feature of the propagation of light as the second postulate of his theory: **The velocity of propagation of light in a vacuum is independent of the relative velocity between the source and the observer (principle of the constancy of the velocity of light).**

The two above-mentioned postulates adopted by Einstein seem to be *logically* inconsistent with each other. Nevertheless, Einstein retained them, since they represented unshakable generalizations of experience. He showed that the inconsistency is only apparent and due to our unjustified, intuitive idea of "absolute simultaneity." By giving an appropriate operational definition of simultaneity for spatially separated events Einstein was able to remove the logical incompatibility between the two postulates. *This definition of simultaneity together with the principle of relativity and the principle of the constancy of the velocity of light form the foundation of the special theory of relativity. The content of the theory of relativity consists of deductions from these basic premises.*

Although the deductions of this theory are perfectly logical, they seem to contradict our "common sense." For instance, two events occurring at two different places may be simultaneous for one observer, but for a second observer event A may have occurred before event B, and a third observer may find event B to have preceded event A. All observers are equally right. *Simultaneity has no absolute meaning.* It depends on the state of motion of the observer.

For slowly moving bodies the predictions of the theory of relativity agree well with those of Newtonian mechanics, which can be considered as an approximation of Einstein's theory, valid for bodies moving with a speed negligible as compared with the velocity of light. At high speeds, however, there is a striking disagreement between the two theories. Experiments show that the predictions of Einstein's theory are the correct ones. The following are the most important deductions of the special theory of relativity:

1. If an object is in motion relative to an observer, its length is found to be contracted in the direction of motion. The measured length approaches the value zero as the relative speed between observer and object approaches the velocity of light.

2. If a clock is in motion relative to an observer, its rate is found to be retarded. The measured period of the moving clock approaches infinity as the relative speed between object and observer approaches the velocity of light.

3. The classical addition formula for velocities is replaced by equation (40.8*b*). According to this equation the relative velocity between two bodies or between a light signal and a body can never exceed the vacuum velocity of light.

4. A mass is not a constant but changes depending on its speed relative to the observer. It increases with the relative speed, tending to infinity as the relative speed between mass and observer approaches the velocity of light.

5. As a result of the preceding statement no mass can be accelerated to the speed of light. The vacuum speed of light is the limiting speed which no material body or energy signal can exceed.

6. Mass and energy are two different forms of the same physical entity. They can be converted into each other in accordance with the relation (40.14*a*) (mass-energy law).

The special theory of relativity shows how the laws of nature can be formulated so as to retain the same form for all systems which move with a constant velocity relative to each other. Einstein generalized the theory of relativity by finding a way of formulating the laws of nature so that they retain the same form for observers moving arbitrarily relative to each other. Einstein's generalization makes use of an additional fundamental assumption which is known as the **principle of equivalence: A homogeneous gravitational field is physically indistinguishable from and perfectly equivalent to an inertial field generated by a constant acceleration.** A second important step consists of *fusion of space and time into a four-dimensional space-time continuum* called the "world." A point in the world is an "event" characterized by three space coordinates (x, y, z) and a time coordinate. The "events" play the same role in this four-dimensional superspace as points do in the familiar three-dimensional space. The use of a four-dimensional world makes it possible to introduce the concept of a curved three-dimensional space, which plays an important role in the theory.

The general theory of relativity deduces the following statements from its assumptions:

1. The space is curved in the vicinity of a mass. The curvature of the space is equivalent to the gravitational field of the mass.

2. The law of gravitation is a special statement of a generalized law of inertia according to which a mass always moves on a so-called *geodesic* line (the "geodesic" is defined as the line which forms the shortest con-

nection between any two of its points). Far from masses the space is not curved, and a body moves in the field-free space in a straight line at constant speed. In the vicinity of a mass the space is curved, and the generalized inertial motion of a second mass in this curved space is an accelerated motion, which may be rectilinear or curvilinear.

3. The geometry in a gravitational field is non-Euclidean. (For instance, the sum of the angles in a triangle is not exactly 180°, and a straight line is not the shortest connection between two points.)

4. A universe filled with mass and energy is curved. The space of such a universe is finite but unlimited (just as the surface of a sphere is finite but unlimited).

The effects predicted by the general theory of relativity are extremely small in cases accessible to our observation. So far, the following phenomena predicted by the theory have been observed: (1) Light of distant stars is bent after passing very close to the sun. (2) Spectral lines in spectra of heavy stars are shifted toward the red (demonstrating the retardation of "atomic clocks" in a gravitational field). (3) The major axis of the elliptical orbit of the planet Mercury executes a slow rotation.

The theory of relativity subjected the fundamental postulates of physics to a radical revision. *The following basic assumptions,* which were generally accepted before Einstein, *were dropped.*

1. It has a physical sense to speak of the absolute velocity of a body or at least of its velocity relative to the ether which fills the infinite space.

2. The statement "events *A* and *B* occurred simultaneously" has always a definite meaning regardless of the separation of the events in space and the state of motion of the observer.

3. The rate of a clock is an absolute quantity; *i.e.*, it is found to be the same by all observers regardless of their velocity relative to the clock.

4. The linear dimensions and the shape of an object are absolute; *i.e.*, they are found to be the same by all observers regardless of their velocity relative to the body.

5. The relative speed between two bodies can assume any value from 0 to ∞.

6. The mass of a body is an absolute constant.

7. Our practical Euclidean geometry is valid near a heavy mass as well as in interstellar space.

8. A clock will run just as fast near a heavy mass as in interstellar space.

9. Inertia and gravity are two independent properties of matter.

Stimulus for a New Theory

In the preceding chapters we made extensive use of the "mass-energy law" originally introduced as an empirical law which could have been established from experience by the process of induction. Actually, however, this law was enunciated by Einstein long before any observations were known which demonstrate the conversion of energy into mass and vice versa. This law was *deduced* by Einstein from the basic assumptions of his *theory of relativity* and has been designated by him as the most important fruit of that theory. Had Einstein not discovered anything besides this law, his name would still be listed among the great contributors to human knowledge. But the mass-energy law is only one of many striking consequences of Einstein's theory which must be considered as one of the most revolutionary feats in human intellectual history.

The events which led to the formulation of the theory of relativity consisted of the discovery of certain phenomena which could not be reconciled with contemporary physical theories and, in fact, appeared to be logically inconsistent with each other. These discoveries were made in the course of experiments which were designed to promote the understanding of the process of propagation of light through the ether.

It appeared that an optical analogue to the following acoustical experiment should yield a positive result. Imagine two observers located on the deck of a long ship. Observer A is stationed at the rear and observer B at the front of the ship. Exactly midway between them is a gun which is fired. Let us assume ideal observers who are able, by means of a stop watch, to determine with unlimited precision the time of arrival of the sound of the shot. When the ship is at rest and the air is still, the sound will arrive simultaneously at A and B. But if the ship moves rapidly forward, the observer B "flees" from the sound wave which is emitted from the gun whereas the observer A moves to meet the pulse wave which travels in his direction. Thus we expect the observer A to stop his stop watch a little before the observer B. [This result is due to the fact that the motion of the sound source (gun) has no effect upon the velocity of the wave which it emits. Once the air has been disturbed, the wave proceeds with a speed which depends only on the pressure and density of the air.] Thus it appears possible to measure the velocity of the ship relative to air. Naturally, the same result would have been obtained were the ship stationary relative to the earth while a wind was blowing from the front toward the rear of the ship. Such measurements reveal only the *relative* velocity between air and ship. It is also important to bear in mind that the same experiment carried out in the enclosed interior of the ship would fail to reveal its motion.

It seems reasonable to expect that it should be possible to demonstrate the motion of our planet relative to the ether through optical experiments analogous in principle to the above acoustical experiment. The earth moves in its orbit around the sun with a speed of about $18\frac{1}{2}$ mi/sec and, in addition, shares the velocity of the entire solar system, which moves rapidly among the stars. But even though this velocity is reasonably large as compared with the speed of light ($\sim$0.01 per cent of c), the success of an optical experiment of this kind depends on our ability to detect time differences in the order of magnitude of 10^{-16} sec! This experimental feat was accomplished in an ingenious fashion by Michelson with his "interferometer" by reducing the time measurement to a measurement of displacement.

Figure 40.1 shows a schematic diagram of the Michelson interferometer (omitting all dispensable details). A parallel beam of light emanating from the source S hits a glass plate P, the surface of which is coated with a thin layer of silver, so that the impinging light is divided into two nearly equally intense beams, one of which is reflected toward mirror M_1 and another one which is transmitted toward mirror M_2. The distances of the two mirrors (M_1 and M_2) from the plate P are nearly equal. After reflection from the mirror M_1, the first beam of light passes through the plate P and enters the telescope T. The second beam is reflected from M_2 back toward the plate P, which reflects it into the telescope (dashed rays in Fig. 40.1).

The two beams of light which thus enter the telescope are coherent, since they have been obtained by division of the light emanating from the same source S. They will therefore interfere with each other. If the mirrors M_1 and M_2 are *exactly* perpendicular to each other, the center of the field of view in the telescope will be uniformly illuminated. If the two beams happen to be exactly in phase opposition, the field of view of the telescope will be dark. When the phase difference between the two beams of light changes (at the point where they merge into a common beam which proceeds toward the telescope), the field of view becomes brighter. Such a change of phase difference can be produced, for instance, by moving mirror M_1 slowly "upward" (in the diagram). This lengthens the path and hence the travel time of the "vertical" beam and hence shifts the phase of the light wave which arrives at P after reflection from M_1. Thus the slow recession of M_1 is accompanied by brightening of the telescopic field of view and a subsequent darkening as the phase difference approaches 180°, etc.

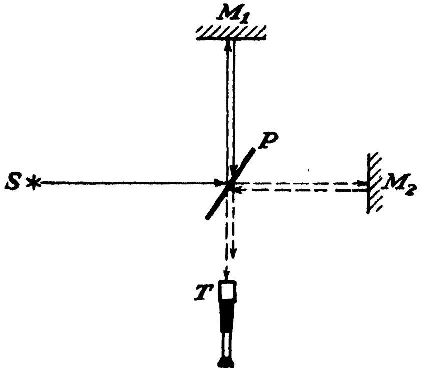


FIG. 40.1. Scheme of the Michelson-Morley experiment.

It is more convenient to perform interferometric observations with the angle between the mirrors deviating slightly from 90°. In this case the field of view is not uniformly bright but rather traversed by parallel dark and bright fringes (similar to the ones shown in Fig. 29.17). As mirror M_1 is moved away from plate P, the fringes are seen to travel sideways. From the observed amount of sideways deviation of the fringes, exceedingly small displacements of M_1 can be determined quantitatively.

The sideways deviation of the fringes seen in the telescope T when the distance between M_1 and P is altered is due to the variation in time which light has to spend traveling from P to M_1 and back. A simple calculation shows that the time required for this round trip can also be changed without altering the distance between P and M_1, namely, when the whole apparatus is moved in the direction from P toward M_1. This round trip lasts a little longer than when the apparatus is at rest. The time required for the light to travel from P to M_2 and back to P is also prolonged by the above motion but not quite so much as for the beam of light which is parallel to the direction of motion of the apparatus. As a result, motion of the apparatus in the direction from P toward M_1 produces a phase shift between the vertical light beam arriving at P from M_1 and the horizontal light beam arriving at P from M_2. Thus if we were to set the apparatus in motion, its motion *relative to the ether* would become apparent, according to our expectations, by a shift in the interference fringes. The velocity could be determined from the magnitude of the shift. Naturally we cannot stop and then start the motion of the earth in order to measure the shift in the fringe pattern, but we can obtain a similar shift by simply rotating the apparatus through 90° so as to interchange the directions of the PM_1 axis and the PM_2 axis.

Using light of 5,000 A and an effective distance (using multiple reflection) of 11 m between M and P, Michelson and Morley expected a maximum shift of one-quarter of a fringe as a result of the rotation of the instrument. They found, however, no such effect. They found a shift of less than one-twentieth of a fringe, which was attributed to experimental error.

The negative result of the Michelson-Morley experiment is inconsistent with the hypothesis that the earth has a relative velocity with respect to the ether. Michelson concluded from this (as was held before him by Stokes) that the ether is dragged along by the earth with its own speed. Thus the Michelson-Morley experiment was, according to this view, analogous to the attempt to detect the motion of the ship in our acoustical analogy by experiments performed in the interior of the ship, where no relative motion between air and observers can exist.

Michelson's conclusion was, however, contradicted by other observations. Observations of Arago and experiments of Fizeau appeared to show clearly that the ether is dragged along with material bodies with velocities appreciably smaller than that of the moving body, so that an "ether wind" relative to the moving body would be expected. Fizeau demonstrated the "ether drag" by sending light through a moving column of liquid in the direction of motion and measuring the increase in the velocity of light. Oliver Lodge, on the other hand, demonstrated that the velocity of light passing near the surface of very rapidly moving bodies is not altered at all, which appeared to show that a moving body does not drag with it an "atmosphere of ether."

The contradictory conclusions about the behavior of the ether to which the abovementioned experiments have led suggest the obvious idea that the ether theory is wrong, *i.e.*, inconsistent with experience, and that it should be abandoned. Heinrich Hertz had already recognized that a luminiferous or an electromagnetic ether endowed with quasi-material properties was a superfluous idea as far as the electromagnetic theory of light was concerned. He realized that the theory needed no more than a set of equations stating how the electromagnetic field varies in space and time. We shall thus eliminate the contradictions which have arisen by abandoning the hypothesis which caused them. This, however, still leaves the observed phenomena unexplained, and we must make new attempts to interpret them.

It appears that the result of the Michelson-Morley experiment could be explained easily on the basis of a particle theory of light. Its negative result would then follow from a general principle which has long been known and which we shall call the *mechanical principle of relativity.* It states: *All mechanical processes proceed according to the same laws in all systems of reference which move with a constant velocity relative to each other.* This means that we use exactly the same equations to solve a dynamic problem whether we are confronted with it inside a uniformly moving train, in a laboratory on firm ground, or in an elevator descending or ascending at constant speed. This statement implies that *it is impossible to detect the motion of a system moving with constant velocity by mechanical observations performed within that system,* since all mechanical events would proceed in it exactly as in another system moving relative to it with an arbitrary constant velocity. Since we do not assume the existence of an ether with respect to which the velocities of all systems could be measured, *the concept of velocity retains a meaning only as a relative velocity with respect to an arbitrary reference system.*

The fact that the experiments of Michelson failed to reveal the motion of the earth relative to its celestial neighbors—suggests that the principle of relativity may be also valid for optics. Subsequent optical and electrical experiments aiming to detect the motion of the earth relative to the stars led to the conviction that *it is impossible to detect the uniform linear motion of a system A relative to a system of reference B by any physical experiments performed within the closed system A.* This led to the following generalization of the **principle of relativity: All physical processes proceed according to the same laws in all systems of reference which move with a constant velocity relative to each other.**

In particular the laws of propagation of light are the same in systems which move relative to each other with constant velocity. For example, a flash of light emanating from a point source will spread out in spherically symmetrical fashion in all directions regardless of whether it was emitted on firm ground or in a fast-moving train. It appears evident on the basis of the particle theory that the corpuscle emitted from the source on the train in the direction of the train motion should have a speed $s_1 = c + v$ relative to ground, where c is the speed of light relative to the train and v is the speed of the train relative to the ground. The speed of the light corpuscles emitted in the opposite direction should be $s_2 = c - v$. This prediction deduced from the corpuscular hypothesis, which explains the result of the Michelson-Morley experiment, is, however, in contradiction to experience. Observations show that the velocity of light coming from a source which moves toward an observer with the velocity v is not $c + v$ but c, in other words, **the velocity of light is independent of the velocity of the source relative to the observer.** This law of nature has been designated by Einstein as the **law of the constancy of the velocity of light.**†

This law is logically inconsistent with the principle of relativity. This inconsistency can be recognized most easily by considering the following concrete example: Imagine that a point source emits light on a train which moves eastward with the velocity v. *According to the relativity principle* the light should (and does) proceed with equal speed relative to the train in all directions, just as a light flash emitted on firm ground would give rise to a spherical wave front. But if this is the case, an observer on firm ground *deduces logically* that the light emitted on the train travels eastward with the speed $c + v$ relative to ground and westward with the speed $c - v$. This is, however, not what he observes. *According to the law of constancy of the velocity of light* he finds the same velocity for the light emitted from the train as does the observer on the train. In other words, he finds that the speed of light emitted from the train westward as well as eastward is equal to c with respect to ground.

How can this contradiction be resolved? Both principles we used are correct, and there is nothing wrong with the logic of the argument!

---

† Among the observations which established this law are those of de Sitter, who studied double stars which rotate about each other so that each of the stars moves periodically toward us and away from us. If the velocity of light were to vary with the relative velocity of the star, certain irregularities in the motion of the stars should be observed. For instance, in certain positions, light emitted simultaneously by the two stars should reach the earth at markedly different times owing to the difference in velocity of the light emitted by them. This should happen when one star moves toward us while the other recedes. As a result, one would, for instance, expect to see at times the same star at more than one place and other peculiar phenomena. But nothing of that sort has been observed. The actual observations confirm the assumption that the velocity of light is not influenced by the motion of the source.

Einstein's Postulates of the Special Relativity Theory

In order to resolve the conflict Einstein adopted (1) *the principle of relativity* and (2) *the principle of the constancy of the velocity of light* as two irrefutable fundamental laws of physics and rejected the validity of the equation $s = c + v$ for the computation of the speed of light, where s is the speed of light relative to ground, c the speed of light relative to the source, and v the speed of the source relative to ground. He rejected this equation because he was able to show that its derivation implied a tacit assumption about the meaning of simultaneity of two spatially separated events which is inconsistent with the law of the constancy of the velocity of light and with the principle of relativity. He substituted for it a new addition theorem for velocities [see equation (40.8*b*)], which he deduced from his basic assumptions and which removes the inconsistency referred to above.

Einstein's step was tantamount to a renunciation of the mechanics of Newton which makes use of the concepts of absolute motion, absolute space, absolute time, and absolute simultaneity. Newton's ideas on space and time are illustrated by the following excerpts from his "Principia": "Absolute motion is the translation of a body from one *absolute place* to another *absolute place*." And "*Absolute*, true and mathematical time, of itself, and by its own nature, flows uniformly on, without regard to anything external." In Einstein's mechanics motion has no *absolute* meaning. Only *relative* velocities between bodies have an operational meaning. This feature was readily accepted by physicists, but the concept of "absolute time" seemed so self-evident and indispensable that many of them put up a desperate fight against Einstein, who claimed that it had no more meaning than "absolute space." "Absolute time" could be thought of as determined by readings of some "world clock" which fixes the time of an event everywhere in the universe for all observers. Einstein showed that such a universal clock is unthinkable, since the time rate of any given clock may be judged quite differently by different observers depending on their state of motion relative to the clock.

At the root of Einstein's modification of mechanics is his analysis of the concept of time, and the pivotal point of that analysis is the analysis of the concept of **simultaneity. Einstein's theory of relativity consists of logical consequences of his definition of simultaneity and of his acceptance of the two basic postulates:**

1. **All laws of physics are the same in all inertial systems** (*relativity principle*).
2. **The speed of light relative to any body is independent of the speed of the source of light relative to that body** (*principle of the constancy of the velocity of light*).

(We define as *inertial systems* all systems which move at constant velocity relative to each other and in which a body upon which no force is acting is not accelerated.)

We shall show in the following section how the judgment of simultaneity of two events *occurring at different points in space* is not an absolute one. It will be seen how two events which are judged by one observer to coincide in time are, with equal justification, ruled by another observer to be separated by a certain time interval.

Relativity of Simultaneity

Let us consider an example used by Einstein and by Bergmann. A train moves past a railroad station. At a certain moment lightning strikes *simultaneously* at two different points of the train *as judged by an observer at the station.* Does an observer located on the train also observe these two lightning strokes as simultaneous events?

To answer this question we must, first of all, define what we mean by simultaneity of two spatially separated events. **We shall define two events occurring at points**

A and B as being simultaneous if the light signals emitted from A and B at the instant of the occurrence of the events reach an observer located midway between A and B at the same instant. *Since the velocity of light is independent of the velocity of the observer relative to the source, this definition is applicable regardless of the state of motion of the observer.*

Figure 40.2 shows the train at the moment when it is struck by the flashes of lightning. The marks left by the lightning strokes on the train and on the station platform are designated by letters a and b and letters A and B, respectively. We could establish by means of a meter stick the exact mid-point M between the marks A and B and station at this point an observer on the platform equipped with a device enabling him to verify the coincidence of the light signals emitted by the lightning flashes at the points A and B. Let us suppose that in our case the observer at M finds the lightning strokes to be simultaneous.

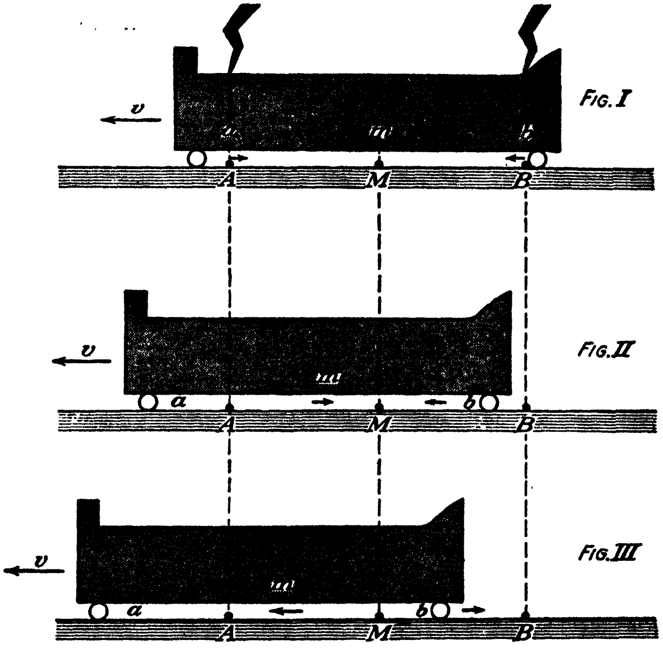


FIG. 40.2.

Since the velocity of the two light signals will have the same value when determined by a moving observer on the train, the same method for checking the simultaneity of the lightning strokes can be used by an observer on the train stationed at m, the mid-point between the marks a and b. It is easy to see, however, that he will not judge the lightning strokes to have occurred at the same time. In Fig. 40.2 I the light signals leave the points at which the marks (A and a) and (B and b) have been produced by the lightning. The light signals are indicated by horizontal arrows under the train. Since the train moves to the left, the light signal coming from the left (Fig. 40.2 II) will reach m before the signal coming from the right (Fig. 40.2 III). Hence, the moving observer will conclude that the lightning struck at a before it struck at b.

It can be similarly shown that, if the moving observer had recorded the two lightning strokes as simultaneous, they would not have been judged to have occurred at the same time by the observer on the platform. **Thus simultaneity is a relative concept;** two events which are simultaneous for one observer may not be simultaneous for another observer who moves uniformly with respect to the first one. In fact, the time sequence of *causally unconnected events* can be reversed according to the state of motion of the observer. Thus, for instance, for observers moving with one (uniform) velocity, the two flashes of lightning may have struck simultaneously; for another one, the lightning at A may have struck first; and for a different one, the flash at B may have preceded the flash at A. *Therefore, when stating that two events are simultaneous, one must specify relative to what observer.*

Relativity of simultaneity has an important consequence in regard to measurements of time and space intervals. Whenever we compare the beats (which we shall imagine to be periodic light signals) of two widely separated clocks which are in relative motion, we must make a judgment of simultaneity. We arrange for the light signals flashed by the two clocks to coincide at a given moment and determine after how many unit intervals another coincidence occurs. It will be shown below that, as a result of the relativity of simultaneity, the speed of a clock is judged to be slower by an observer who moves relative to the clock than by an observer relative to whom the clock is at rest.

When we measure a moving length, the judgment of simultaneity enters our operation in the following fashion: Let us suppose we scratch a scale between the marks A and B on the station platform in Fig. 40.2 and use it as a meter stick to measure the distance a-b between the marks on the moving train. There is no difficulty in measuring a-b while the train is at rest. We simply place it over our "meter stick" and read the scale. But how can we measure the distance a-b when the train is moving? We can take two *simultaneous* scale readings by determining with what mark on our scale the point b coincides at the moment when point a coincides with A. As will be shown below, as a result of the relativity of simultaneity, the moving length will be found to be contracted.

The Lorentz Transformation

According to the principle of relativity all laws of physics should have the same form in all inertial coordinate systems. In particular, the velocity of light should have the same value in all such systems. There is no difficulty in showing that the laws of mechanics, such as the law of inertia or the law of conservation of momentum, remain the same, whether we express them in terms of coordinates at rest with respect to ground or in terms of a coordinate system moving with respect to ground. If x, y, z are the coordinates of the system S which is fixed to ground, we have merely to replace them by the coordinates x', y', z' of the system S' which is moving with constant velocity relative to ground in order to obtain an expression of the law in the new system. This coordinate transformation does not change the type of equation expressing the law. The relationship between the coordinates of the systems S and S' (*Galilean transformation*) can be read off from Fig. 40.3. The system S' (x', y', z') is moving with velocity v with respect to the system S (x, y, z) parallel to the X axis. The coordinates of the point P can be stated in either system. The relationship between the primed and unprimed coordinates follows from the diagram:

$$\begin{aligned} x' &= x - vt \\ y' &= y \\ z' &= z \\ t' &= t \end{aligned} \qquad \textbf{(Galilean transformation)} \qquad (40.1)$$

Self-evident as this transformation seems to be, it has to be abandoned, since it proves to be inconsistent with the fact that the velocity of light coming from a given light source has the same value regardless of whether it is measured by an observer in the system S or by one moving with the system S′. The breakdown of the Galilean transformation is due to the fact that it is based on the tacit assumptions that (1) measurements of time and (2) measurements of length are unaffected by the state of motion of the coordinate system from which they are judged. We have mentioned above that these assumptions are untenable as a consequence of the relativity of simultaneity.

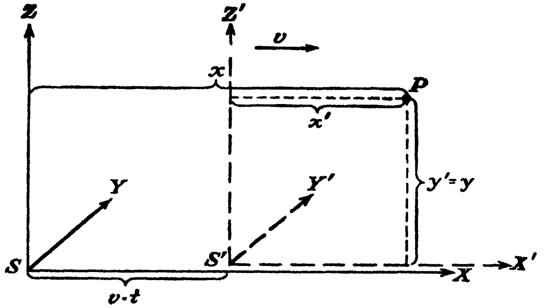


FIG. 40.3.

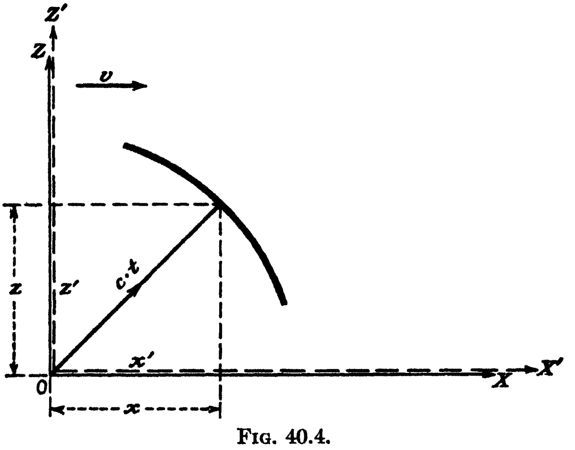


FIG. 40.4.

In order to derive a coordinate transformation which would be consistent with Einstein's postulates, let us imagine the origins of the two coordinate systems of Fig. 40.3 to coincide at the point 0 of Fig. 40.4. At the moment of coincidence of the two origins we set the clocks at the origins of S and S' at 0. A source of light at 0 emits a light wave which covers in t seconds the distance ct. The two observers in the systems S and S' will make the following statements about the coordinates of the points reached by the light signal:†

---

† In Fig. 40.4 $y = 0$ for points located in the plane of the paper. For such points this equation degenerates into the Pythagorean theorem $x^2 + y^2 = (ct)^2$.

$$x^2 + y^2 + z^2 = (ct)^2$$

and

$$x'^2 + y'^2 + z'^2 = (ct')^2$$

which is the same as

$$x^2 + y^2 + z^2 - (ct)^2 = 0 \qquad (a)$$

and

$$x'^2 + y'^2 + z'^2 - (ct')^2 = 0 \qquad (b)$$

It is now easy to see that $x^2 + y^2 + z^2 - (ct)^2 \neq x'^2 + y'^2 + z'^2 - (ct')^2$ if we substitute the values of t', x', and z' from the Galilean transformation (40.1). *The law of constancy of the velocity of light demands, however, that the two sides of the above expression be equal;* in other words, the following equation must hold:

$$x^2 + y^2 + z^2 - (ct)^2 = 0 = x'^2 + y'^2 + z'^2 - (ct')^2 \qquad (40.2)$$

Einstein showed that this equation is valid if one uses for the transition from the unprimed coordinates to the primed coordinates, instead of the Galilean transformation, a transformation of the following form:

$$\begin{aligned} x' &= k(x - vt) \\ y' &= y \\ z' &= z \\ t' &= p(t - qx) \end{aligned} \qquad (40.3)$$

where k, p, and q are functions of the relative velocity between the systems S and S'. Substituting (40.3) into (40.2) and equating the coefficients of corresponding terms, simple algebra yields the following values for k, p, and q:

$$\begin{aligned} k = p &= \frac{1}{\sqrt{1 - (v^2/c^2)}} \\ q &= \frac{v}{c^2} \end{aligned} \qquad (40.4)$$

Substituting these values in (40.3), we obtain the transformation sought by Einstein:

$$\boxed{\begin{aligned} (a)\ x' &= \frac{x - vt}{\sqrt{1 - (v^2/c^2)}} \\ (b)\ y' &= y \\ (c)\ z' &= z \\ (d)\ t' &= \frac{t - (v/c^2)x}{\sqrt{1 - (v^2/c^2)}} \end{aligned}} \quad \textbf{(Lorentz transformation)} \quad (40.5)$$

This group of equations is known under the name of *Lorentz transformation,* since they were proposed originally by Lorentz in an *ad hoc* explanation of the Michelson-Morley experiment. They were a mathematical device devoid of tenable physical justification. *Einstein, on the other hand, deduced this set of equations from the invariance of the speed of light and the principle of relativity.* The striking feature of these equations is that they imply the transformation not only of the coordinate in the direction of motion (x') but also of the time coordinate t'. The kinematic implications of this transformation will be discussed in the following section.

RELATIVISTIC KINEMATICS

Contraction of Lengths

Suppose we have a rigid rod of length L_0 *at rest in the system S'* of Fig. 40.3. Let us assume that this rod is placed along the X' axis, its left end being at the origin;

The coordinate of the right end of the rod will then be $x' = L_0$. At the moment when the origin of the system S' (and hence also the left edge of our rod) passes the origin of the coordinate system S, we attempt to measure the length of the rod by means of a meter stick which is *at rest in the system S*. What value do we get for the length of the moving rod?

We choose the time $t = 0$ as the moment when the origins of the two systems coincide and take a simultaneous reading (simultaneous as judged by the observer in S) of the position of the right end of the rod on our scale. The reading L, which we should get, is obtained from equation (a) of the Lorentz transformation (40.5) by setting $t = 0$, $x' = L_0$, and $x = L$; we obtain

$$x' = L_0 = \frac{L}{\sqrt{1 - (v^2/c^2)}} \tag{40.6}$$

or

$$\boxed{L = L_0 \sqrt{1 - \frac{v^2}{c^2}}} \tag{40.6a}$$

We see that the moving rod appears contracted.

Equation (40.6a) shows that there is no such thing as the "real" length of an object. Its length *in the direction of motion* depends on the relative velocity of the system from which the object is observed. The object appears to have a maximum length to an observer who is at rest with respect to it. In this case $v = 0$, and hence $L = L_0$. We see from the same equation that an object would be seen to shrink to zero length in the direction of its motion by an observer relative to whom the object moved with the velocity of light. In this case $v = c$, and

$$L = L_0 \sqrt{1 - (c^2/c^2)} = L_0 \sqrt{1 - 1} = 0.$$

Thus *the speed of light appears, for the first time, in the role of a limiting speed.* It is unreasonable to assume that an object could move faster, since then $v > c$ and $v^2/c^2 > 1$. The expression under the square root would become negative, and the length of the object would "contract" to an imaginary thickness which has no physical meaning.

The same result is obtained by assuming the rigid rod to be fixed in the system S and the meter stick to be at rest in the system S'.

Dilatation of Time

Let us now consider how the progress of time will be judged by an observer who moves relative to a clock. Assume a clock at the origin of the system S' of Fig. 40.3. It will have permanently the coordinate $x' = 0$ in that system.

Consider the time interval of 1 sec between two consecutive beats of the clock pendulum which mark the time readings $t_1' = 0$ and $t_2' = 1$ sec in the S' system. Substituting the value of $x' = 0$ in equation (a) of the Lorentz transformation (40.5) we obtain

$$0 = \frac{x - vt}{\sqrt{1 - (v^2/c^2)}}; \qquad \text{hence } x = vt \tag{1}$$

Substituting equation (1) and the value $t' = 0$ in equation (40.5d) we get

$$0 = \frac{t - (v/c^2)vt}{\sqrt{1 - (v^2/c^2)}} = \frac{t[1 - (v^2/c^2)]}{\sqrt{1 - (v^2/c^2)}} = t\sqrt{1 - (v^2/c^2)}$$

hence

$$\underline{t = 0}$$

Now we substitute $x' = 0$ and $t_2' = 1$ sec in the equations (*a*) and (*d*) of the Lorentz transformation. For $x' = 0$ in (*a*) we get, as above in (1),

$$x = vt \tag{1a}$$

Substituting the value of x from equation (1*a*) and the value of $t' = 1$ sec in equation (40.5*d*), we get

$$1 = \frac{t - (v/c^2)vt}{\sqrt{1 - (v^2/c^2)}} = \frac{t[1 - (v^2/c^2)]}{\sqrt{1 - (v^2/c^2)}} = t\sqrt{1 - \frac{v^2}{c^2}}$$

hence

$$\underline{t = \frac{1}{\sqrt{1 - (v^2/c^2)}}}$$

Thus we see that the second beat is perceived by the observer in the S system, not after 1 sec (as it is by the observer in the S' system), but after a longer time interval,

$$t = \frac{1}{\sqrt{1 - (v^2/c^2)}} \text{ sec}$$

The observer in the S system finds that the clock is slow, whereas the clock keeps "correct" time with respect to the observer in the S′ system who is at rest with respect to the clock.

In general, any time interval which lasts t_0 seconds in the S' system is measured to last

$$\boxed{t = \frac{t_0}{\sqrt{1 - (v^2/c^2)}}} \tag{40.7}$$

seconds in the S system which is in relative motion with respect to the S' system. *The same result is obtained by exchanging the position of the clock and of the observer, i.e.*, by placing the former in S and the latter in S'.

Thus, we see that a clock in a given system runs fastest for an observer who is at rest with respect to it ($v = 0$) and runs more slowly for an observer who moves with respect to it.

An observer moving with the velocity of light ($v = c$) *with respect to a reference system observes the clock in that system to stand still.* Any time interval lasting t_0 seconds in the system S will be judged to last $t = \frac{t_0}{\sqrt{1 - (c^2/c^2)}} = \frac{t_0}{\sqrt{1 - 1}} = \infty$, *i.e.*, infinitely long, by an observer in a system S' which moves with the velocity of light relative to the system S. *We see again the limiting character of the velocity of light;* for $v > c$, $v^2/c^2 > 1$ and the square root becomes imaginary.

By comparing equations (40.6*a*) *and* (40.7) *we see that the factor by which a length is contracted as a result of relative motion is reciprocal to the factor by which a time interval is prolonged.*

Addition of Velocities

Let us imagine, for example, that the point P of Fig. 40.3 is moving with the velocity u' with respect to the system S', which, in turn, moves parallel to the X axis with the velocity v relative to the system S. How large is the velocity of P relative to the system S? According to Galilean kinematics this question is easily answered: The velocity u of P relative to S is given by the sum $u = v + u'$, where v is the relative velocity between the systems S and S' and u' is the velocity of P

relative to S'. If the point P should move with the velocity of light with respect to S', $u' = c$ and $u = v + c > c$.

If we use the Lorentz transformation to answer our question, we shall find a different result. Let us divide equation (a) of (40.5) by equation (d), writing u' for x'/t', where u' is the velocity of a point relative to the S' system, whereas $x/t = u$ is the velocity of the same point relative to the S system. We get

$$u' = \frac{x'}{t'} = \frac{(x - vt)/\sqrt{\cancel{1 - (v^2/c^2)}}}{[t - (v/c^2)\,x]/\sqrt{\cancel{1 - (v^2/c^2)}}} = \frac{x - vt}{t - (v/c^2)x} = u' \tag{40.8}$$

Dividing numerator and denominator by t we get

$$u' = \frac{(x/t) - v}{1 - (v/c^2)(x/t)} = \frac{u - v}{1 - (vu/c^2)} \tag{40.8a}$$

hence,

$$u - v = u'\left(1 - \frac{vu}{c^2}\right) = u' - u\frac{vu'}{c^2}$$

Rearranging terms we get

$$u' + v = u\left(1 + \frac{vu'}{c^2}\right)$$

Solving for u we obtain **the relativistic formula for the addition of velocities:**

$$\boxed{u = \frac{v + u'}{1 + (vu'/c^2)}} \tag{40.8b}$$

We see that the equation (40.8b) for the addition of velocities, derived with the aid of the Lorentz transformation, differs from the classical formula $u = v + u'$ by a factor $1\Big/\left(1 + \frac{vu'}{c^2}\right)$.

This formula has the interesting property that it makes it impossible for the velocity of light to be exceeded relative to any observer. This can be seen from the following *examples*:

1. **Consider $v = \frac{3}{4}c$ and $u' = \frac{3}{4}c$.** According to the classical formula

$$u = v + u' = \tfrac{3}{4}c + \tfrac{3}{4}c = 1.5c$$

that is, $u > c$. But according to Einstein's formula, we get

$$u = \frac{\frac{3}{4}c + \frac{3}{4}c}{1 + \dfrac{\frac{3}{4}c \times \frac{3}{4}c}{c^2}} = \frac{1.5c}{1 + \dfrac{9c^2}{16c^2}} = \frac{1.5c}{\frac{25}{16}} = \frac{16 \times (1.5)c}{25} = \frac{4.8}{5}c < c$$

Thus we get the surprising result that, if a passenger were riding a bicycle with three-quarters of the velocity of light on a train which itself is moving with respect to ground with three-quarters of the velocity of light, his speed with respect to ground would still be smaller than the velocity of light.

2. **Consider $u' = c$ and $v = c/2$.** u' is the velocity of a light signal moving with the velocity $u' = c$ relative to S', which itself moves with the speed $v = c/2$ relative to S.

We obtain from equation (40.8*b*)

$$u = \frac{(c/2) + c}{1 + \frac{(c/2)c}{c^2}} = \frac{1.5c}{1 + \frac{1c^2}{2c^2}} = \frac{1.5c}{1.5} = c$$

The light signal moves with the same speed *c* relative to both systems!

3. **Consider $u' = c$ and $v = c$.** From equation (40.8*b*) follows

$$u = \frac{c + c}{1 + (cc/c^2)} = \frac{2c}{1 + 1} = c$$

This is the same speed as obtained in Example 2 under different conditions.

We see from equation (40.8*b*) *that in the case of velocities encountered in ordinary practice, where the product $u'v$ is negligibly small as compared with c^2, the classical formula $u = v + u'$ is an adequate approximation.*

RELATIVISTIC DYNAMICS

Variability of Mass

After showing how the length of a rigid rod and the time interval between the beats of a clock pendulum assume different values for an observer who is in motion relative to them from those for an observer who is at rest with respect to the rod or the clock, we shall show how the mass of a body is found to change, depending on its state of motion relative to the observer.

Let us consider the impact of two balls from the standpoint of an observer in the system S' of Fig. 40.3. He chooses two balls of equal mass m' and imparts to them equal and opposite velocities in the direction of the X' axis so that they move toward each other with the velocities u' and $-u'$ relative to the S' coordinate system. The two balls are perfectly *inelastic.* They collide and stop, forming a body of mass $2m'$.

Let us now consider this process from the standpoint of the observer in the system S, with respect to which the system S' moves with velocity v. We shall not commit ourselves to the assumption that the two balls will necessarily have the same mass as judged by the observer in the system S. The observer in the sytem S will call the masses of the two balls m_1 and m_2.

The velocities of the balls relative to the observer in the system S before the impact follow from the equation (40.8*b*).

$$u_1 = \frac{v + u'}{1 + (vu'/c^2)} \qquad (40.9a)$$

and

$$u_2 = \frac{v + (-u')}{1 + [v(-u')/c^2]} = \frac{v - u'}{1 - (vu'/c^2)} \qquad (40.9b)$$

After the impact the two balls fuse into one body, which remains at rest in the S' system; hence, its velocity relative to the S system is v. The law of conservation of momentum will be expressed by the observer in the system S as follows:

$$m_1u_1 + m_2u_2 = (m_1 + m_2)v \qquad (40.10)$$

After substituting the values of u_1 and u_2 from equations (40.9*a*) and (40.9*b*) into (40.10) and rearranging terms, we obtain

$$m_1\left(1 - \frac{u'v}{c^2}\right) = m_2\left(1 + \frac{u'v}{c^2}\right) \qquad (40.11)$$

We now eliminate v from this equation by finding an expression for $u'v/c^2$ in terms of u_1 and u_2 from equations (40.9*a*) and (40.9*b*). The algebraic computation, which is omitted here, yields

$$\frac{u'v}{c^2} = \frac{2c^2 - u_1^2 - u_2^2 - 2\sqrt{(c^2 - u_1^2)(c^2 - u_2^2)}}{u_1^2 - u_2^2}$$

Substituting this expression for $u'v/c^2$ in (40.11) results in the simple expression

$$m_1\sqrt{1 - \frac{u_1^2}{c^2}} = m_2\sqrt{1 - \frac{u_2^2}{c^2}} \tag{40.12}$$

In order to realize the physical implications of this equation most easily, we shall assume that the motion of the second ball before the collision was such that it happened to be at rest in the system S; *i.e.*, we assume $u_2 = 0$.

Then equation (40.12) becomes

$$m_1 = \frac{m_2}{\sqrt{1 - (u_1^2/c^2)}} \tag{40.12a}$$

What does this equation tell us? It tells that $m_1 > m_2$, since $\sqrt{1 - (u_1^2/c^2)} < 1$ as long as $|u_1| > 0$. When $u_1 = 0$, $m_1 = m_2$. *In other words, the two balls are judged by the observer in the system S to have equal masses as long as they are both at rest in that system* ($u_1 = u_2 = 0$). *If, however, one of the bodies* (m_1) *is moving with the velocity u_1, its mass is observed to increase.*

If, instead of using the symbols m_2 for the mass at rest and m_1 for the moving mass in equation (40.12*a*), we use the customary symbols m_0 for the *rest mass* and m for the moving mass, we can write the **formula for the relativistic mass:**

$$\boxed{m = \frac{m_0}{\sqrt{1 - (v^2/c^2)}}} \tag{40.12b}$$

where *v stands now for the relative velocity between mass and observer* [v corresponds to u_1 of equation (40.12*a*)].

This result shows that **the mass of a body is not a constant** which determines its inertia, *i.e.*, its tendency to resist acceleration. **It appears rather as a function of the velocity** of the body relative to the reference system with respect to which the velocity and the mass are determined.

Equation (40.12*b*) *illustrates most clearly the character of the velocity of light as a limiting velocity.* We see that m increases as the velocity v grows. Finally, when $v = c$, we get $m = \dfrac{m_0}{\sqrt{1 - (c^2/c^2)}} = \dfrac{m_0}{0} = \infty$. This means **that the mass of a body tends to infinity as the body approaches the velocity of light.** This means that the body is rendered less and less accelerable as its velocity approaches that of light. It is thus impossible to accelerate the body above the speed of light, since its mass becomes infinite and the body becomes unaccelerable when v reaches the values of c. **Thus we realize that the velocity of light is the highest velocity which a material body can reach.**

The Mass-Energy Law

It can be shown mathematically that the equation $1/\sqrt{1 - \Delta^2} = 1 + \frac{1}{2}\Delta^2$ holds approximately if Δ^2 is very small compared with 1. Let us apply this approximation to equation (40.12*b*), assuming $v \ll c$. We obtain

$$m = m_0\left[1 + \frac{1}{2}\left(\frac{v}{c}\right)^2\right] = m_0 + \frac{m_0}{2} v^2 \frac{1}{c^2}$$

hence,

$$mc^2 = m_0c^2 + \frac{m_0}{2} v^2$$

or

$$\boxed{\frac{m_0}{2} v^2 = (m - m_0)c^2} \tag{40.13}$$

$m_0v^2/2$ is the expression of Newtonian mechanics for the kinetic energy of a body. We see that, according to Einstein, this energy is equal to the increase in mass $(m - m_0)$ multiplied by the square of the velocity of light.

By writing the equation as follows,

$$m = m_0 + \frac{m_0}{2} \frac{v^2}{c^2} = m_0 + \frac{E}{c^2} \tag{40.13a}$$

we can also interpret it as follows: The increase in mass of a moving body over the value of the rest mass m_0 is due to a term E/c^2 (kinetic energy/c^2) which can be interpreted as a mass increase due to energy gain. [The terms mc^2 and m_0c^2 of equations (40.13) and (40.13*a*) have the dimension of energy (the same as $(m_0/2)v^2$), and the term E/c^2 has the dimension of mass.]

These equations suggest a fundamental generalization of the law of conservation of energy, namely, a combination of the laws of conservation of mass and of conservation of energy into one single law. Einstein was able to show that any kind of energy (*e.g.*, heat, potential energy, radiant energy, etc.) **possesses inertia and that the mass corresponding to the energy E is given by**

$$\boxed{m = \frac{E}{c^2}} \tag{40.14}$$

He concluded that mass could be converted into energy such that the amount of energy liberated when the mass m disappears is given by

$$\boxed{E = mc^2} \tag{40.14a}$$

Since mass can be converted into energy and vice versa, we are justified in considering mass as a form of energy or, better, mass and energy as two different forms of the same entity.

Equation (40.13) can be written

$$mc^2 = m_0c^2 + \frac{m_0}{2} v^2$$

We can interpret it as follows: Any mass has a certain **rest energy** m_0c^2. When the mass is moving the energy of the mass is mc^2. It exceeds the rest energy by the amount of the kinetic energy acquired.

Relativistic Paradoxes

It is not difficult to understand why the theory of relativity encountered tremendous opposition before it was finally generally accepted. Its consequences are seemingly strikingly contradictory to our "common sense." But actually, there is no contradiction to our common sense. On the contrary, the theory is based on common-sense reasoning. The discrepancy lies between some of its predictions and our common *experience*. The reason for this discrepancy is easy to see. If we put $v = 0$ in all relativistic formulas, we obtain formulas of classical mechanics. *The discrepancies arise only when we make predictions about the behavior of bodies which move with velocities approaching the speed of light.* Since we do not commonly observe such rapid bodies, their behavior is beyond the range of our everyday experience, and hence it is not surprising that their behavior sounds very strange to us.†

Let us review some of the paradoxes predicted by Einstein's theory in order to appreciate the strain to which it seems to put our common sense.

1. You and I are moving past each other with a high speed. Each of us is holding a meter stick parallel to the direction of motion. You measure my meter stick, and you find that it is contracted to the length of 70 cm. I measure yours at the same time and find that *yours* is contracted to the length of 70 cm. This does not seem to make sense! If both of them are contracted to 70 cm, should we not both find them to be equal?! If we can't both be right, who is right? This situation is mildly confusing!

2. Two babies are born in two neighboring villages. The birth of each of the babies is celebrated by a pistol shot at the exact moment of the birth. The pilots of three very rapidly moving rocket planes observe the light flashes of the shots. The pilot A concludes that the babies were born simultaneously. The pilot of plane B reaches the conclusion that the baby of village V_1 is the younger one, and the pilot of plane C decides that the baby of village V_2 is the younger one of the two babies. Which of the babies is "really" younger? The disturbing thing is that, according to the theory of relativity, such a question has sense only if we specify in what system of reference the seniority of the babies is to be decided. All three pilots are right. There is no logical contradiction between their statements, although to a layman they seem to disagree!

3. You know that, if you ride a bicycle at the speed of 30 mi/hr in the direction of motion on a train which moves relative to ground at 60 mi/hr, your speed relative to ground will be 90 mi/hr. But when you move with a speed of 2.9×10^{10} cm/sec in the direction of motion on a train which is moving with the speed of 2.9×10^{10} cm/sec with respect to ground, your velocity relative to ground will be less than 3×10^{10} cm/sec!

4. If you are on a very fast-moving train and flash a light signal which spreads in all directions, you will find the velocity of light to be the same with respect to the train whether you measure it in the direction of motion of the train or in the opposite direction or in any other direction. An observer on firm ground studying the same flash of light will get the same results as you.

5. Imagine yourself to be the son of a youthful, adventurous scientist who takes off in a rocket ship to explore the universe soon after you are born. Shortly after the take-off he reaches a velocity not much smaller than the velocity of light and continues moving through interstellar space with nearly constant speed. His space

---

† In modern "nuclear engineering," electrons and ions moving at speeds approaching the velocity of light are now commonly used. Thus the relativistic behavior of such particles has entered the realm of everyday experience of nuclear physicists.

ship is equipped with a powerful television transmitter and a receiver so that he can remain in communication with the earth during his trip. He also has a "calendar clock," which he sets correctly before his departure.

You are brought regularly to the television station where you can see your father on the screen and from where your picture can be broadcast to him. Your father bores you terribly. He makes faces to amuse you, but he goes about it so slowly! His quickest smile may last for 5 min. It is particularly tedious to watch him eat; it may take him 20 hr to get through with breakfast. His pace of life seems to be very leisurely indeed. But soon you learn to read the "calendar clock," and one day you make a revealing discovery: You find that, whereas you have aged 10 years since your father's departure, his calendar clock has advanced little more than 1 month. Thus, though your father was very sluggish according to your clock, his speed was normal according to his own clock, which you find almost 100 times as slow as yours.

Your father does not seem to age at all. As you look at him (by this time you are a dignified old gentleman with a white beard in your ninetieth year of age), he looks just like the photograph taken before his departure. Your father's behavior is somewhat annoying to you. It is irritating to see him still trying to amuse you as if you were a two-year-old baby. But he can't be blamed. *Your calendar clock was found by him to be just as slow as his clock appeared to you.* You did not seem to him to have aged any more than he appeared to you to have aged.

According to the theory of relativity the same observations as described above would be made, regardless of whether the space ship moved away from the earth, toward the earth, or in any other direction.

Experimental Verifications of the Theory

The value of a theory is assessed partly by the number of known phenomena which it is capable of explaining and partly by its fruitfulness in predicting new phenomena. Einstein's special theory of relativity was eminently successful in both respects. We can mention only a few of the many experimental verifications of the theory.

1. The theory predicts the existence of the Doppler effect and yields the correct formula for the frequency shift.

2. The time dilatation in a moving system was very nicely demonstrated recently by observations of the lifetime of mesons. As we mentioned in Chap. 38, mesons decay with emission of an electron. The mean "lifetime" of mesons is the average time which elapses between the moment when we observe a meson and the moment when we observe its disintegration. This lifetime is a characteristic constant for mesons at rest, corresponding to the lifetime of a radioactive element for a particular nuclear species. According to the theory of relativity this lifetime (like any other time interval) should appear longer to an observer relative to whom the mesons are in rapid motion. This means that fast-moving mesons should decay at a slower rate than slow-moving mesons. This has been actually verified by observers of cosmic rays.

3. The well-known alteration of the velocity of light which passes through a liquid by the streaming of the liquid "with the light" or "against the light" is predicted by the theory. The calculated "drag coefficient" agrees excellently with the observed value. The effect is explained simply by the relativistic addition theorem for velocities [equation (40.8*b*)] without resort to any hypotheses about the ether.

4. In Chap. 38 we mentioned the fact that the mass of beta ray electrons has been found to vary with speed. Exact measurements showed that this variation of mass with velocity is in precise accordance with equation (40.12*b*).

5. The processes of "annihilation" of positron-electron pairs and of "pair production" by conversion of a γ ray energy quantum into a positron and an electron demonstrate the possibility of complete conversion of mass into energy and vice versa.

6. The exact equality (in nuclear transformations) between the loss of mass and the gain in kinetic energy of emitted nucleons and recoil nuclei is a brilliant confirmation of Einstein's mass-energy law and thus also of the theory of relativity from which this law was deduced.

Need for a More General Theory

According to the special theory of relativity, general laws of nature formulated in terms of one coordinate system S do not change their form when they are expressed relative to another coordinate system S' which moves with a constant velocity relative to S, the coordinates of the S' system being correlated with the coordinates of the S system according to the Lorentz transformation. The situation radically changes, however, when we extend our considerations to systems which are accelerated with respect to each other. Then one cannot say any longer that the laws of nature retain the same form for all arbitrarily moving systems. Consider, for instance, the system S (Fig. 40.3) to be the inertial system which we arbitrarily label as our "rest system." Suppose the system S' is represented by a box which is accelerated uniformly in the direction of the X axis relative to the S system. It is easy to see that, whereas the law of inertia holds in the S system, it does not hold in the S' system. A body upon which no forces are acting will move with a constant speed relative to system S but will have an acceleration relative to the system S'.

This example may make one doubt whether a way of formulating physical laws could be found which would preserve the validity of the laws for observers moving with arbitrary nonconstant velocities relative to each other. Such a general method of formulating physical laws was, however, eventually developed by Einstein. This method consists of a formulation of the laws in terms of so-called Gaussian coordinates and of suitable rules of transformation from one system of Gaussian coordinates to another. This method of description enabled Einstein to formulate his new **principle of general relativity** as follows:

All laws of nature have the same form (for different systems in any state of motion relative to each other) when expressed in Gaussian coordinates.

The general laws of physics are expressed in Einstein's general theory of relativity in such a way that only *relative* positions, velocities, and accelerations among bodies enter the equations.

It is beyond the scope of this text to present the details of this difficult, highly abstract mathematical theory. We shall merely consider briefly some of its physical aspects without discussing Gaussian coordinates. Let us consider, as an example of denial of a preferential status to any coordinate system in the general theory of relativity, the inertial forces observed in a rotating system. If we spin, for instance, a liquid in its container, centrifugal d'Alembert forces appear which tend to move the liquid away from the center of rotation, thus deforming the liquid surface into a paraboloid (Fig. 40.5). This experiment seems to indicate clearly the possibility of detecting *absolute* rotational motion in space. But are we really sure that the liquid surface would be deformed as shown in the figure in an isolated pail of water spinning in an empty universe? The statement "an isolated body is spinning in an empty universe" has obviously no kinematic operational meaning; we have to assume at least a second body with reference to which the rotation of the first body could be determined. But if we have a second body, how does the observer stationed on the first body know whether it is the first body that spins or the second body that revolves about the first body? He would see the same relative motion in both cases! One is tempted to reply to this: "It will not be hard for him to make a decision, since, if the first body is spinning, the observer on that body will detect centrifugal inertial forces, whereas in the second case (revolution of body 2 about body 1) he will not." But we have actually no strong experimental justification for such a statement. It is supported by evidence such as the experiment of Newton in which he rotated a pail around its water content (the water remaining initially at rest) watching for a possible deformation of the surface of the water before it began to participate in the rotation of the pail. He found no effect. But it was suggested by Mach that, if a very much larger mass (a container with a much heavier wall) were used instead of the pail, an effect might have been observed. According to Mach we have no experimental evidence to justify a refutation of the statement that the water surface in the pail of Fig. 40.5 would be deformed into a paraboloid if the whole universe were to revolve about it. Nobody ever tried such an experiment!

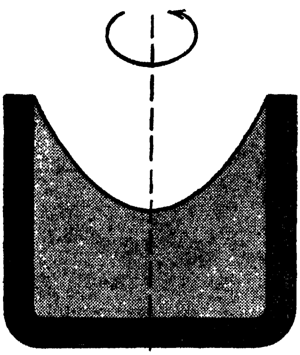


FIG. 40.5. Deformation of the surface of a rotating liquid.

In Einstein's general theory of relativity neither the rotating body nor the surrounding universe is considered as a preferred system of reference. The statement that the pail rotates with respect to the fixed stars is equivalent to the statement that all the stars revolve about the pail; no physical experiment could decide which is "really" the case. In other words, the Ptolemaic theory is considered just as correct as the Copernican theory. The latter has merely the advantage of greater simplicity of description.

Principle of Equivalence

The assumption that centrifugal inertial forces would be observed on the surface of a sphere if the rest of the masses of the universe were to revolve about it has an interesting implication. It makes the manifestations of inertia dependent on other masses. So far it was only the gravitational force which was considered as an interaction between a mass and other masses, whereas we spoke of inertia (*i.e.*, tendency to resist acceleration) of an individual body without suspecting that it may be, like gravity, a consequence of the presence of other masses.

There is actually a fundamental relationship between inertia and gravity which is suggested by our common experience. We know that in a gravitational field all bodies experience the same acceleration irrespective of their mass: $a = F/m =$ constant. This implies that the *gravitational mass of a body to which F is proportional is proportional to the inert mass m.* Eötvös has confirmed this proportionality by careful experiments with a precision of 1 part in 10 million. The proportionality between inert and gravitational mass was known before Einstein, but it was considered as mere coincidence without deeper significance. *Einstein saw in this proportionality the revelation of a basic law of nature and utilized it as the empirical foundation of his general theory of relativity.*

Einstein realized that a distinction between gravitational and inertial forces cannot be upheld operationally. *The impossibility of distinguishing between inertial forces* *in an accelerated system and gravitational forces in a homogenous gravitational field, in other words, the complete equivalence of manifestations of inertia and gravity, is a fundamental assumption of Einstein's general theory of relativity. It is known as* **the principle of equivalence.**

The principle of equivalence is illustrated by the imaginary experiments which have been discussed in Chap. 6. The observer in the accelerated box (Fig. 6.3) will observe that all objects fall toward the right with the same acceleration $(-a)$ regardless of mass. Every mass will also seem to possess a weight of $-ma$. In fact, if he cannot look out of his box, he will not be able to decide by any experiment whether what he is observing are consequences of an acceleration of his box relative to the stars or simply a revelation of the presence of a very large mass to the right of his box which gives rise to a gravitational field inside his box. In an accelerated elevator (Figs. 6.1 and 6.2) we can consider that the quasi-gravitational field caused by the acceleration of the elevator is superimposed upon the field of gravity. Thus the gravitational field is observed to diminish inside the elevator when it is accelerated downward. In fact, the field disappears for the observer inside the elevator if the latter is allowed to fall freely. *We see thus that a gravitational field may be judged to exist or not to exist in a certain region according to the state of motion of the observer.* Whereas a gravitational field will be judged to be the same by all observers in different inertial systems, the magnitude (and direction) of a given gravitational field will be judged differently by observers who are accelerated relative to each other. *Thus we see how gravity enters considerations designed to generalize the theory of relativity by extending its validity to systems accelerated relative to each other. These considerations suggest that a general theory of relativity may lead to a theory of gravitation and that the law of inertia and the law of gravitation may fuse into one law. This was actually the end product of Einstein's work.*

Gravity and Non-Euclidean Geometry

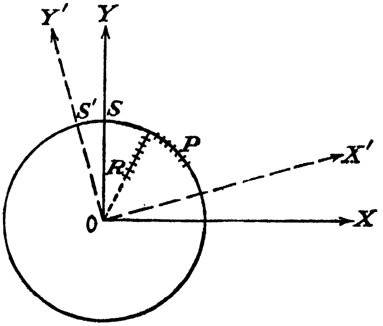


FIG. 40.6.

Let us now see how the equivalence of inertial and gravitational forces suggests a new, radical modification of our notions of space and time. Imagine a flat merry-go-round (Fig. 40.6) on which we draw an X' and a Y' axis and call this rotating disk our S' system. The axes X and Y drawn on firm ground will define our nonrotating system of reference (system S). We know that, if we were to measure the periphery of a circle and then to divide it by the measured value of its diameter, we should get the number $\pi = 3.14159 \ldots$ Let us now imagine that the observer in the system S' (which we assume to rotate very rapidly) decides to determine π. Before he goes to work, we ask him to place a pin at one point of the periphery of his merry-go-round so as to scratch a circular groove on the ground identical in diameter with the diameter of his merry-go-round. To determine π he will lay off tiny measuring sticks along the periphery of the rotating disk as indicated at P, and then he will measure the diameter by aligning his measuring rods as shown at R. But according to the special theory of relativity his measuring rods in positions P should be shortened (in comparison with the rods laid off radially) as they are laid off tangentially and are thus aligned in the direction of motion.† As the observer on the merry-go-round lays off his meter sticks along the circumference of the merry-go-round, he measures by the same operation the circumference scratched by the pin into the ground. Owing to the contraction of his measuring rods, he will get a smaller value of π than the value obtained by Archimedes. Owing to the contraction of the measuring rods, which depends on their orientation, we expect the geometry on the disk to be quite different from the plane geometry we learned from Euclid. *We expect the geometry on the rotating disk to be non-Euclidean.*

In addition to the distortion of the geometry on the disk, we shall also expect the time scale to vary from place to place. A clock placed at the origin O will keep time like a clock in the system S, but a clock on the periphery of the disk should go slower owing to the great peripheral speed. The slowing of time will depend on how far the clock is from the center of the disk. As the observer in the S' system starts on a trip from the center toward the periphery of the disk, we shall observe in the system S how his geometry (for instance, the value of π) and his time scale gradually change. At the same time the observer in S' will observe changes in centrifugal, quasi-gravitational, inertial forces (an increase toward the rim) as he approaches the rim. This association between the changes in geometry and time scale with a change in the inertial quasi-gravitational force field suggests the following question: "*Can we conclude on the basis of the principle of equivalence that similar changes would take place for an observer moving from regions of lower gravitational field strength into regions of greater gravitational field strength?* Should we expect a change in geometry and a slowing of clocks as we approach a heavy star?" Einstein answered this question in the affirmative. *According to the general theory of relativity, meter sticks contract and clocks (as well as all processes, such as atomic frequencies and physiological processes) slow down according to the intensity of the gravitational field to which they are exposed.*

The Four-dimensional World

Once again Einstein has forced us to modify our notions of space and time. In addition to the contraction of rigid rods and slowing of clocks which manifest themselves to a uniformly moving observer, we have to accept the modifications of space and time scale in gravitational fields. But the reform of our notions of space and time, which sprang from the theory of relativity, is even more radical than has been indicated so far. Minkowski suggested treating time on an equal basis with the space coordinates x, y, z. This amounts to the construction of a four-dimensional "superspace," which is normally referred to as the *Minkowski world.* This is a rather natural step which introduces great mathematical elegance into the theory of relativity and without which the development of the general theory would probably have been impossible.

Whereas in geometry we are concerned with the location of points in space which can be defined by the three cartesian coordinates x, y, z that locate the event in space and the time coordinate which determines the "moment of its occurrence," in the four-dimensional Minkowski "world" such an event is a "world point" determined by the four coordinates x, y, z, t. In other words, Minkowski fused space and time into a four-dimensional "space-time."

† The application of the special theory of relativity to processes on the periphery of a rotating disk is permissible as long as we consider only small lengths moving for a very short time interval during which the velocity can be assumed not to change appreciably.

Einstein's Theory of Gravitation

Minkowski's approach transformed kinematics into geometry. A point moving in space at constant velocity is represented by a straight line in the four-dimensional Minkowski world. As we know, this type of motion (uniform linear motion) is possible only in regions far removed from matter, *i.e.*, in the absence of gravitational fields. *According to Einstein the four-dimensional "superspace" as well as the space of three dimensions (the portion of the "world" which we can visualize) is curved in the vicinity of masses.* As a result, a mass does not move uniformly in a straight line in the vicinity of another mass but rather moves, in general, on a curve with acceleration. *In Einstein's theory of gravitation the motion of a mass in a field of gravity is looked upon as "inertial motion" in a curved "world."*

Verifications of the General Theory

The modification of "space-time" in the vicinity of masses affects not only the path of matter but also the path of light. If the observer in an accelerated box throws a stone at right angles to the direction of the acceleration of his box (*i.e.*, "horizontally"), he will observe the stone to move in a parabola, just as in a homogeneous field of gravity. For the same reasons, we should expect the path of photons in a horizontal light beam to move in a parabola. But of course, the parabola will be much less curved in this case because the photons are so much faster than the stone. Now *according to the principle of equivalence, the same should be expected to happen in a gravitational field.* A beam of light should be curved similarly to the path of a horizontally projected stone. This prediction of Einstein's theory was verified quantitatively in 1919. An astronomical expedition observed the position of stars whose light was passing very close to the edge of the sun during a solar eclipse. It was demonstrated that light was actually curved near the sun (the path being concave toward the sun) by the amount predicted by Einstein's theory.

Another brilliant verification of Einstein's theory of gravitation was the explanation of the precession of the perihelion of the planet Mercury. The orbit of this planet is an ellipse whose perihelion (point of the orbit nearest the sun) revolves slowly around the sun. This observable motion of the perihelion (precession) could not be entirely accounted for by the perturbation of Mercury in its orbit by attractive forces of other planets. A precession of 43 sec of arc per century (!) remained unaccounted for. Einstein's theory of gravitation, however, predicted this rotation of the axis of the elliptical orbit by exactly the observed amount.

The third effect predicted by Einstein is the "gravitational red shift." Owing to the dilatation of the time scale in a gravitational field, light coming from a very heavy star should appear to us redder (longer period of vibration) than the corresponding light (for instance, the light of the sodium D line) coming from a terrestrial source. This effect is very difficult to ascertain because the same effect (red shift) could be also produced by motion of the star away from us. It has been, however, verified beyond doubt by astronomical observations on the double-star system of Sirius (see Prob. 17).

Cosmology

When we stand in an open field on a dark clear night and peer into the universe, it seems to us in our poetic mood that this world is infinite in extent and eternal in time. But Newtonian mechanics imposes a prosaic limitation on our dreams. An infinite universe is inconsistent with Newtonian mechanics. The gravitational field intensity at the surface of a sphere is given by $E = G(M/R^2)$, where G is the gravitational constant, R the radius of the sphere, and M the mass within the sphere. If the average density of matter in the universe is d, we can express the field intensity at the surface of a sphere which we imagine described with radius R about an arbitrary point as follows:

$$E = G\frac{M}{R^2} = G\frac{4/3\pi R^3}{R^2}d = G\frac{4}{3}\pi Rd$$

We see that the field increases with the radius. And if the world is infinite, we can always consider any point as being on the surface of a sphere whose center is infinitely far away. The field intensity at the point should thus be infinite. But a world where the gravitational field intensity is infinite everywhere is not a very satisfactory choice, and so we discard it.

We shall try as our next assumption a world which has a center of maximum star density. In this world the stars and matter get less and less abundant as we recede from the center; this universe can be considered as an oasis in infinite empty space. Einstein did not like this world either. Such a world could not be eternal. Energy emitted in form of light and other radiations would wander off to infinity, and the world would eventually die of exhaustion of energy.

The non-Euclidean geometry of curved "space-time" suggested to Einstein a "better world." It is a world which is *not infinite* in extent but, rather, *unlimited.* This statement does not seem to make sense. How can a space be finite if it has no limits? Don't we need a definite boundary to limit the extent of a space? The finite unlimited space can be understood with the help of an analogy. Let us consider the surface of a sphere which is coated with a *very thin* layer of water inhabited by protozoa. These flat creatures will have no idea of the existence of anything beyond the film of water which they inhabit. Their "world" will be the surface of a sphere. It will be finite in area but unlimited; *i.e.*, a protozoan could move forever in any direction without ever encountering "the end of his world." *Even though the spherical shell on which they live is an object in a three-dimensional world, they will have no idea of what a third dimension means.* All their addresses are perfectly determined by two coordinates, the geographic longitude and latitude. Their thinking of space will be entirely in terms of two dimensions, since their *ordinary* daily experience never reveals the existence of a third dimension. Since they are very tiny in comparison with the radius of the globe and since their activities and measurements are confined to small areas, their geometry will be practically Euclidean; *i.e.*, the sum of the angles in their tiny triangles will be 180°, parallel lines will not meet on their drawing boards, and π will have the value of 3.14159 . . . . But let us suppose that a unicellular genius undertakes to geometrize on a large scale. He will draw a triangle† much larger than the world has ever seen before. What will he discover? Something we know very well from spherical geometry: The sum of the angles in a spherical triangle is larger than 180°. He will draw large circles and find a π of less than 3.14159 . . . . These discoveries will inspire in him a bold idea. He will say good-by to his friends and depart, moving in a straight line (a great circle). Much to the surprise of his friends, who gave him up as missing in action, he will finally reappear at the same spot, arriving from the opposite direction. He will publish then a revolutionary and nearly unintelligible paper. He will say: "Our world is not two-dimensional, but it is a *two-dimensional curved surface in a three-dimensional*

---

† What he calls straight lines will be shortest connections between two points. Such lines are arcs of so-called *great circles*. A great circle is obtained by the intersection of the spherical surface with a plane which passes through the center of the sphere.

space. We cannot visualize the third dimension, but we can infer its existence from the experiments discussed in my paper."

Now let us match the mental feat of the protozoan, and let us imagine our space to be a *three-dimensional curved space in a four-dimensional world.* We cannot imagine a four-dimensional space-time, and a curved space has no meaning to us in terms of a visualizable image. But like the protozoa, we could observe phenomena analogous to what they saw, which would reveal to us the curvature of our space. For instance, we could measure the sum of the angles in huge triangles in space and see if it is 180°. Gauss did it. He chose three mountain tips in Germany as the corners of the triangle. He found no deviation from 180°. But this is no proof that our space is "flat"! His triangle may not have been big enough.

What else could we try to demonstrate the curvature of our universe?—If 10 protozoa had left the north pole of their sphere, moving in various directions along great circles (meridians), they would eventually have been "focused" at the south pole; *i.e.*, they would have met there and passed each other and eventually come back to the north pole. If our space is curved, we should be able to send 10 people or, better, 10 rays of light in various directions in space, expecting them to come back eventually to the point of origin (coming from the opposite direction) after having been focused in the meantime at the opposite pole of the universe. If we remember that many stars are so far away from us that light coming from them takes millions of years to reach us, we realize how large our universe must be and do not expect anybody to perform such an experiment and to wait patiently for the return of the light beams.

But if the decision on whether or not we live in a curved universe is so hopeless, what is the point of speculating about such a possibility? First of all the curvature of our universe is a necessary consequence of Einstein's theory of gravitation. Einstein deduced that any space which contains matter must be curved. Second, this kind of a universe will not be in danger of losing its energy content by leakage to infinity and hence will be a suitable model of an eternal universe. Attempts have been made to estimate the radius of such a "static curved universe" from determinations of the average density of matter in the universe. One estimate yields $R \simeq 3 \times 10^{11}$ light years; *i.e.*, light will need 3×10^{11} years to traverse its radius.

But Einstein's static universe proved to be unstable and had to be given up. Observations seem to indicate that we live in an **expanding universe.** We mentioned in Chap. 31 the red Doppler shift of distant spiral nebulae. It seems that they all move away from us at great speed; the farther away they are the faster they move; *i.e.*, the velocity of recession is proportional to their distance from us. We have the impression of being at the center point from which they flee. What is so formidable about us to cause this general flight of the spiral nebulae from us?

This observation can be explained as a harmless illusion which an observer would have no matter where he was standing in the universe. The explanation can be based on the assumption of an expansion of the plain, ordinary Euclidean space. Imagine a checkerboard with a figure at the center of each field. Imagine now the board to expand in such a fashion that the relative velocity between two neighboring figures is the same everywhere. Consider now the figures along one row, say figure 5 of row 1. Suppose that the distance between figures 5 and 6 expands at the rate of 10 cm/sec. Then the relative speed between figures 6 and 7 will also be 10 cm/sec, and hence the relative speed between figures 5 and 7 will be 20 cm/sec. Thus it is easy to see that figure 5 will have the impression that figure 7 moves away from it faster than figure 6. In general, figure 5 will conclude that figures 6, 7, 8, etc., move faster away from it the farther away they are, the velocity of recession being proportional to distance. But looking backward at figure 4, figure 5 will have the same impression! The distance between figures 5 and 4 grows at the rate of 10 cm/sec and between figures 5 and 3 at the rate of 20 cm/sec, etc. Thus figure 5 will find that all figures flee from it along *all* directions, receding with a speed proportional to their distance from it.

Although this simple explanation based on the assumption of an expanding Euclidean space is adequate to explain the geocentric illusion of the recession of spiral nebulae, it cannot be reconciled with certain observations on the distribution of distant spiral nebulae in space. A tolerable agreement is obtained by assuming an expanding, curved, three-dimensional space in a four-dimensional "space-time."

Again, we have to resort to our analogy with a world of fewer dimensions than ours in order to explain the preceding statement. Imagine a two-dimensional universe in a three-dimensional world. The two-dimensional universe may be imagined as a rubber balloon with stars painted on its surface. When we blow it up, the stars recede from each other as the surface grows. An observer stationed at any of these stars will have the illusion of being in a central position from which all stars recede symmetrically. Similarly, our impression of the recession of spiral nebulae from us can be interpreted as an illusion to be expected in a three-dimensional curved universe expanding in a four-dimensional world.

If the student feels slightly dizzy at this point, he should not be discouraged. This is the price one often has to pay for climbing to the summits of human knowledge.

QUESTIONS AND PROBLEMS

1. Why did Einstein drop the ether hypothesis of light propagation?

2. What is the empirical foundation of the special theory of relativity?

3. How did Einstein explain the result of the Michelson-Morley experiment?

4. State Einstein's postulates of special relativity.

5. What assumptions of the Newtonian mechanics did Einstein drop?

6. Explain relativity of simultaneity.

7. Suppose you are watching an observer in a fast-moving train. Will his clock appear to you to be slow? Will it appear to him to be slow? Will your clock appear to him to be slow?

8. How fast must a train move relative to you in order to appear to be contracted to half its length.

9. Do you expect an observer to find a change in the frequency of the sodium *D* line in the spectrum of a distant star which is moving at high speed at right angles to the line joining the star to the observer?

10. How fast must an electron move in order to appear as massive as a proton?

11. Why cannot a material body exceed the speed of light?

12. From what considerations did Einstein infer the equivalence of mass and energy?

13. What does the "principle of equivalence" state?

14. What is your conception of "curved space"?

15. Can the sequence of two causally connected events appear reversed to two observers moving at constant speed relative to each other? For instance, can an observer hold that a man was shot before the trigger of the murder weapon was pulled whereas another observer asserts that the man was shot after the trigger was pulled?

16. Two shots are fired *independently*. Could two observers disagree on their time sequence?

17. Two stars of a double-star system can be resolved in a telescope. One of the star masses is very much larger than the other. What observations and computations would you perform to verify the Einstein red shift?

18. Imagine a space ship moving with a speed of 10^{10} cm/sec. In this ship there is a clock which emits pulses of radio waves at intervals of 1 sec as judged by an observer on the earth. What would be the length of this time interval if that clock had been removed from the space ship and placed on the earth?

19. Find the mass of an electron moving with the speed of 2×10^{10} cm/sec.

20. Describe *quantitatively* how a spherical particle of $R = 10^{-8}$ cm radius, moving with a speed of 2×10^{10} cm/sec relative to the observer, will appear in shape. Will it be a sphere? How large will it be? Will it remain symmetrical?

21. How fast do you have to move a source of light in order to double the velocity of the light which it emits in a vacuum?

22. Imagine a train moving with the speed of $v = 10^{10}$ cm/sec relative to ground. How fast must you move relative to the train in the direction of motion in order to reach half the speed of light relative to ground?

23. How fast must one move radium atoms in order to double their half life?

24. How large is the rest energy of an electron in ergs?

TABLE III. RELATIONS BETWEEN UNITS†

Length

1 meter (m) = 100 centimeters (cm)
1 micron (μ) = 10^{-4} cm
1 millimicron (mμ) = 10^{-7} cm
1 angstrom unit (A) = 10^{-8} cm
1 inch (in.) = 2.540 cm
1 foot (ft) = 30.48 cm
1 mile (mi) = 5,280 ft
1 mile = 1,609 m

Mass

1 kilogram (kg) = 1,000 grams (gm)
1 kilogram = 2.205 pounds (lb)
1 slug = 32.2 lb

Velocity

1 mi/hr = 44.7 cm/sec
60 mi/hr = 88 ft/sec
1 ft/sec = 30.48 cm/sec

Force

1 lb = 4.45×10^5 dynes
1 gm-force = 981 dynes

Pressure

1 atmosphere (atm) = 14.7 lb/in.^{2} = 1,013,000 dynes/cm^2
1 lb/in.^{2} = 69,870 dynes/cm^2

Energy

1 ft-lb = 1,356 joules = 0.3239 calorie (cal)
1 joule = 10^7 ergs = 0.239 cal
1 cal = 4.186 joules = 3.087 ft-lb

Power

1 watt = 10^7 erg/sec = 0.239 cal/sec
1 horsepower (hp) = 746 watts = 550 ft-lb/sec

---

† For relations between electric units, see text, p. 384.

TABLE IV. DENSITIES OF SOME COMMON MATERIALS†

Solids	Density, gm/cm^3	Liquids	Density, gm/cm^3
Aluminum	2.70	Mercury	13.56
Brass	8.44–8.70	Glycerin	1.26
Copper	8.93	Water (at 4°C)	1.00
Lead	11.34	Alcohol	0.79
Platinum	21.50	Olive oil	0.92
Ice	0.917	Gases at 0°C, 76 cm Hg	
Cork	0.22–0.26	Air	0.001293
Pinewood	0.35–0.50	Carbon dioxide	0.001977
Glass	2.4–2.8	Hydrogen	0.000090

---

† The density in the cgs system is numerically equal to the *specific gravity*, which is defined as the ratio of the density of the given substance to the density of water.

TABLE V. APPROXIMATE VALUES OF COEFFICIENTS OF EXPANSION

Substance	Coefficient of linear expansion α (per °C)	Substance	Coefficients of cubical expansion β (per °C)
Aluminum	22×10^{-6}	Mercury	18.2×10^{-5}
Brass	19×10^{-6}	Water (~20°C)	20.7×10^{-5}
Copper	17×10^{-6}	Ethyl alcohol	110×10^{-5}
Pyrex glass	3.2×10^{-6}	Ether	164×10^{-5}
Steel	11×10^{-6}		

TABLE VI. SPECIFIC HEATS

(Approximate values at about 20°C unless specified otherwise)

Substance	*Specific heat, cal/(gm)(deg C)*
Aluminum	0.21
Brass	0.09
Copper	0.092
Glass (crown)	0.16
Ice (at −20–0°C)	0.50
Iron	0.11
Mercury	0.033
Alcohol (ethyl)	0.58
Steam (at atmospheric pressure and temp. 100–200°C)	0.48

TABLE VII. MELTING POINTS AND HEATS OF FUSION

Substance	Melting point, °C	Heat of fusion, cal/gm
Aluminum	658	76.8
Lead	327	5.86
Mercury	−39	2.82
Oxygen	−219	3.30
Water	0	79.71

TABLE VIII. BOILING POINTS (AT ATMOSPHERIC PRESSURE) AND HEATS OF VAPORIZATION

Substance	Boiling point, °C	Heat of vaporization, cal/gm
Water	100	539
Mercury	357	65
Alcohol	78.3	204
Oxygen	−182.9	50.9
Helium	−268.6	6

TABLE IX. INDICES OF REFRACTION

(For the wave length of the yellow sodium *D* line)

Substance	
Diamond	2.417
Carbon disulfide	1.630
Dense flint glass	1.6555
Crown glass	1.5171
Water	1.3330
Air (at 0 °C and 76 cm Hg)	1.000243

TABLE X. SPECIFIC RESISTANCES OF CONDUCTORS

Substance	Specific resistance, ohms/cm at 0°C	Temperature coefficient per °C
Silver	1.63×10^{-6}	0.0038
Copper	1.70×10^{-6}	0.0039
Aluminum	2.83×10^{-6}	0.0045
Tungsten	5.5×10^{-6}	0.0045
Platinum	11×10^{-6}	0.0038
Manganin	48×10^{-6}	0.00001
Carbon	4×10^{-3}	−0.00025

TABLE XI. DIELECTRIC CONSTANTS

Substance	*Dielectric constant K*
Air (at 0 °C and 76 cm Hg)	1.00059
Sulfur	4.0
Glass	6
Mica	5.7
Water	81.1

REFERENCES FOR FURTHER STUDY

General Physics

Margenau, H., W. W. Watson, and C. G. Montgomery, "Physics—Principles and Applications," 1 vol., McGraw-Hill.

Sears, F. W., and M. W. Zemansky, "College Physics," 2 vols., Addison Wesley.

Grimsehl, E., "A Textbook of Physics," 5 vols., Blackie.

Mechanics

Mach, E., "The Science of Mechanics," Open Ct.

Grimsehl, E., "A Textbook of Physics," Vol. 1, Blackie.

Synge, J. L., and B. A. Griffith, "Principles of Mechanics," McGraw-Hill.

Heat and Thermodynamics

Cork, J. M., "Heat," Wiley.

Fermi, E., "Thermodynamics," Prentice-Hall.

Zemansky, M. W., "Heat and Thermodynamics," McGraw-Hill.

Loeb, L. B., "Kinetic Theory of Gases," McGraw-Hill.

Grimsehl, E., "A Textbook of Physics," Vol. 2, Blackie.

Acoustics

Wood, A., "Acoustics," Interscience.

Optics

Valasek, J., "Optics," Wiley.

Monk, G. S., "Light: Principles and Experiments," McGraw-Hill.

Wood, R. W., "Physical Optics," Macmillan.

Jenkins, F. A., and H. E. White, "Fundamentals of Optics," McGraw-Hill.

Electricity and Magnetism

Sydam, V. A., "Fundamentals of Electricity and Electro-magnetism," Van Nostrand.

Müller, R. H., R. L. Garman, and M. E. Droz, "Experimental Electronics," Prentice-Hall.

Reich, H. J., "Theory and Applications of Electron Tubes," McGraw-Hill.

Atomic Physics

Born, M., "Atomic Physics," Stechert.

Semat, H., "Atomic Physics," Rinehart.

Oldenberg, O., "Introduction to Atomic Physics," McGraw-Hill.

Pollard, E., and W. L. Davidson, "Applied Nuclear Physics," Wiley.

Modern Physics

Richtmeyer, F. K., and E. H. Kennard, "Introduction to Modern Physics," McGraw-Hill.

Eddington, A. S., "Space, Time, and Gravitation," Cambridge.

De Broglie, L., "Light and Matter," Dover.

Philosophy of Science

Frank, Ph., "Between Physics and Philosophy," Harvard University Press.

Bridgman, P. W., "The Nature of Physical Theory," Princeton University Press.

Lindsay, R. B., and H. Margenau, "Foundations of Physics," Wiley.

Margenau, H., "The Nature of Physical Reality," McGraw-Hill.
Bavink, B., "The Natural Sciences," Francke.
Cohn, M. R., and E. Nagel, "Introduction to Logic and Scientific Method," Harcourt Brace.

Biophysics

Glasser, O., "Medical Physics," Year Book Publishers, Inc.
Stuhlman, O., "Biophysics," Wiley.
Rashevsky, N., "Mathematical Biophysics," University of Chicago Press.
Schrödinger, E., "What Is Life?" Macmillan.

INDEX

A

Abampere, 365
Aberration, chromatic, 561
  spherical, 546, 561
Abramson, 335
Absolute motion, 13
Absolute temperature, 231, 299
Absolute time, 9, 844
Absolute zero, 230
Absorption, of radiation, 675
  of sound, 468, 469
Absorption coefficient, 675
Absorption spectra, 591
Absorption spectrum, X-ray, 738
Acceleration, angular, 26
  average, 15
  centripetal, 76, 78–81
  definition of, 15, 16, 75
  of gravity, 20
  instantaneous, 16
Accelerators for ions, 802, 826–828
Accidental events, 292
Accommodation of eye, 569
Achromatic systems, 562
Action, concept of, 129
  at distance, 187
  and reaction, 49, 51
Action potentials, 327–329
Adiabatic changes, 260, 261, 279, 280
Air columns, vibrations of, 483, 484
Airplane, 157
Alpha decay, quantum theory of, 791*ff*.
Alpha particles, 777
Alpha-ray spectra, 784*ff*.
Alpha rays, 777
Alternating-current generator, 609*ff*.
Alternating currents, 609*ff*.
  vector representation of, 615*ff*.
Ammeter moving coil, 385, 386
Ampère's molecular currents, 381
Ampere unit, 365
Amplification factor, 354
Amplifier, 354
  gain of, 354*ff*.
Amplitude, 430
amu (*see* Atomic mass unit)
Analyzer, 655
Anderson, 766
Anderson, C. D., 808, 809
Angle, of incidence, 460
  of reflection, 460
  of refraction, 461
Angstrom unit, 520
Angular acceleration, 26
Angular impulse, 96
Angular magnification, 579
Angular momentum, law of conservation of, 97–99
  quantisation of, 726
Angular velocity, 26
Animal electricity, 318, 326
Animal heat, 235–238
Annihilation of electron pairs, 813*ff*.
Annihilation radiation, 813
Anode, 312
Antenna radiation, 661
Antinodes, 478
Arago, 658, 842
Archimedes, 141, 860
  principle of, 141–145
Aristotle, 16
Arrhenius, S., 314
Aspirator, 157
Astigmatism, 570
Astrophysics, 593
Atmospheric pressure, 137, 148, 149
Atomic bomb, 822
Atomic diameter, 728
Atomic distances (*see* Interatomic distances)
Atomic energy, 831*ff*.
Atomic hypothesis, 336
Atomic mass, 770
Atomic mass unit, 770, 804
Atomic model, of Bohr, 725*ff*.
  of Rutherford, 706*ff*., 709, 721–725
  of Schrödinger, 757–759
  of Thomson, 703*ff*.
Atomic number, 689, 769*ff*.
  and nuclear charge, 708
Atomic rays (*see* Molecular rays)
Atomic structure, 733*ff*.
Atomic weights, 735
Atomizer, 157
Atoms, excitation of, 730, 736–739
  and molecules, 277
Attraction, gravitational, 173, 174
  intermolecular, 282–285
Atwood's machine, 46
Audibility, range of, 466
Ausculation, 492
Avogadros law, 267, 300
Avogadro's number, 305

B

Balance, analytical, 38
  torsional, 191
Ballisto-cardiograph, 56
Balmer formula, 701
Balmer series, 701
Band spectrun, 592, 593
Bar magnet, lines of force, 198

Barer, R., 573, 585, 607
Barkla, 677, 679
Barnett, 383, 410
Barnett effect, 383
Barometer, 148
Batteries, 318–323
Beams, J. W., 71
Beats, 486, 487
Becker, 804
Becquerel, H., 772, 774
Bell, 470
Bennett, 577
Bergmann, P. G., 844
Bernoulli, Daniel, 156
Bernoulli effect, 156–159
Bernoulli equation, 154–156
  generalization of, 161
Beta-ray spectrum, 785
Beta rays, 776
Betatron, 418, 419
Bethe, 823
Binding energy, nuclear, 771
Biot's law, 362*ff.*
Biot-Savart law, 361, 362
Bird, flight of, 157, 158
Black-body radiation, 252
Blind spot, 569
Blood-flow measurement, 144, 161, 416
Blood-pressure measurement of, 138*ff.*
Body temperature of animals, 235
Boell, E. J., 142, 150
Bohr, N., 693, 695, 727, 742, 756, 800, 824
Bohr's theory, of heavy atoms, 733*ff.*, 739*ff.*
  of hydrogen atom, 725–731, 739*ff.*
Boiling point, 225
  elevation of, 247
Bolometer, 348
Boltzmann, L., 293
Bomb, atomic, 822
Bombardment, by alpha particles, 799*ff.*
  by neutrons, 815*ff.*
  by protons, 802*ff.*
Boomerang, 159
Born, M., 303, 762
Bothe, 804
Boyle's law, 261
  derivation of, 295–298
  deviations from, 261*ff.*
Bragg, W. L., 676, 685, 690
Bragg's reflection law, 686
Brahe, Tycho, 172
Brain tumors, location of, 331, 834
Brewster's law, 655
Bridgman, P., 8
British system of units, 42
British thermal unit (Btu), 232
Brown, 335
Brownian motion, 286, 306, 307
Btu, 232
Bulk modulus, 456
Bunsen ice calorimeter, 255
Buoyancy, 140, 141
Buoyancy, centripetal, 142
Burch, 573
Burton, 693

C

Calorie, 232
  work equivalent of, 128, 130, 259, 260
Calorimeter, 234
Calorimetry, 233
Camera, photographic, 566–568
  pinhole, 496
Canal rays, 341
Cancer, treatment of, with radioactive materials, 833*ff.*
Candle, international, 511
Candlepower, 511
Capacitance, 210, 211, 215
  of plate condenser, 213–214
  of sphere, 211–213
Capacitive impedance, 623–627
Capacitor, 211
Capacity, electrical, 210, 211
  thermal, 233
Carnot, Sadi, 126, 254
Carrier wave, 665
Cartesian diver, 143
  and microbalance, 143
  and microrespirometer, 142
Cascade jumps, 732*ff.*, 737
Caspersson, T., 607
Cassini, 507
Cataphoresis, 311
Cathode, 312
Cathode-ray oscilloscope, 386, 387
Cathode rays, 341
Cauchy, 659
Causality, 761*ff.*
Cavendish, H., 176, 332
Cells, photoelectric, 719
  voltaic or galvanic (*see* Batteries)
Celsius, (*see* Centigrade scale)
Center, of gravity, 100–101
  of mass, 60, 80, 100
Centigrade scale, 226, 227
Centrifugal force, 81, 82
Centrifuge, 69
Centrifuge microscope, 70
Centripetal acceleration, 76, 78–81
Centripetal force, 81
Cgs system, 42
Chadwick, J., 708, 805*ff.*
Chain reactions, 821–824
Change of state (*see* Phase changes)
Charge, conservation of, 189
  elementary, 338–340
  moving, magnetic field of, 366, 367
    in magnetic field, 371–374, 409, 410
  on surface of conductor, 210
  unit, 191, 365
Charges, like and unlike, 188, 189
Charging, by contact, 208
  by induction, 208

Chromatic aberration, 561
Circuit laws, for alternating current, 630
  for direct current, 346–350
Cloud chamber (*see* Wilson chamber)
Coates, 335
Cockroft, 802
Coefficient, of expansion, 228
  of friction, 123
Coherent light sources, 517
Coil, flat, magnetic field of, 364
  long, magnetic field of, 380
Collisions, elastic, 54–56, 287, 288, 294
  inelastic, 55, 56
Colloidal solutions, 307
Colors, 502–505
  complementary, 505
  of thin films, 529
Combustion, heat of, 234
Compass, gyro, 383
  magnetic, 186
Complementarity, 755*ff*.
Complementary colors, 505
Components of a force, 82
Composition, of forces, 73, 76
  of motions, 73, 74
Compound microscope, 572
Compressibility, of gas, 243, 261
  of liquids, 133
Compton effect, 718, 743*ff*.
Concave mirror, 546–554
Concentration potentials, 325
Concepts, 7, 14
Condensation, heat of, 242
Condensers, 211
  with dielectric, 215
  discharge of, 624
  in parallel, 221
  plate, 215
  in series, 221
  spherical, 214
Condon, 791
Conduction, in electrolytes, 315–316
  in gases, 340
  of heat, 249, 250, 280
  in metals, 342–345
Conductivity, thermal, 249, 250, 344, 345
Conductors, 190
  charge on, 210
  field within, 208*ff*.
Cones and rods in retina, 568
Conic sections, 177
Conservation, of charge, 189
  of energy, 115, 120, 233, 785*ff*.
  of mass, 119
  of momentum, 53–57, 119
  principles of, 54, 55
  of work, 113
Consonance, 491
Contact theory of electric batteries, 319
Continuity equation, 153
Continuous spectrum, 504
Convection, of heat, 250–252
Convection current, 251
Converging lens, 560*ff*.
Conversion table, electrical units, 384
Convex mirror, 554–557
Cooling by evaporation, 248
Cornea, 568
Corpuscular theory of light, 499–502, 505, 506
Corti rods, 492
Cosmic rays, 780
Cosmology, 861*ff*.
Coulomb, C. A., 191
Coulomb unit, 192
Coulomb's law, for charges in vacuum, 190–192
  in dielectric, 192
  for magnetic poles, 192
Counter, Geiger-Müller, 778
  scintillation, 777
Crawford, 235*ff*.
Cream separator, 69
Critical angle, 512
Critical mass, 822
Critical point, 244
Critical pressure, 244
Critical temperature, 244
Crooks, 671
Crystal, as diffraction grating, 681*ff*.
  reflection of X rays from, 685*ff*.
Cubical expansion, 228
Curie, Marie, 773, 819
Curie, Pierre, 773, 819
Current, alternating, 609*ff*.
  chemical effects of, 312–314
  convention on direction of, 316
  direct, 310, 316
  displacement, 649*ff*.
  effective, 613–615
  electric, 310, 311
  force on, in magnetic field, 367–370
  induced, 393–395
  magnetic field of, 360
  measurement of, 385, 386
  polarization, 324
  strength of, definition of, 310
    unit of, 310*ff*., 364, 365
Curvilinear motion, 77
Cyclotron, 826*ff*.

D

D'Alembert forces, 65
  centrifugal, 67
D'Alembert's principle, 64–67
Dalton, J., 277
Damped oscillations, 639
Daniell cell, 320
Dark-field illumination, 576
da Vinci, Leonardo, 106
Davisson, 752
Davy, 319
de Broglie, 740
de Broglie equation, 750

de Broglie waves (*see* Matter waves)
Debye and Scherrer method, 690
Decay constant, 790
Decibel, 470
Deduction, 7, 20, 21
Definitions, operational, 8, 9, 13
Dempster, A. J., 375
Density, 37, 147
Depth measurements, sonic, 472
Descartes, R., 142, 187, 695
de Sitter, 843
Detector circuit, 666
Deuterium, 376, 816
Deuteron, 828
Dewar flask (*see* Thermos bottle)
Diamagnetic media, 193
Diamagnetism, 220, 421
Diameters, atomic, 728
  molecular, 291, 301–304
  nuclear, 707*ff*.
Diathermy, 640
Dielectric, field strength in, 216
  polarization of, 216–219, 649
  surface charge on, 217
Dielectric constant, 192
Dielectric materials, 216*ff*.
Differential equations, 362
Differential laws, 362
Diffraction, circular-hole, 528
  of electrons, 751*ff*.
  grating, 521–524
  by single slit, 524–527
  of X rays, 681*ff*.
Diffusion, 278
Dimensions of physical quantities, 44
Diode, 666
Diopter, 559
Dipole oscillator, 641
Dipoles, electric, 216
  magnetic, 185
Dirac, P. A. M., 381, 740, 810
Dirac's positron theory, 810*ff*.
Directional hearing, 492
Discharge, electric, through gases, 340–342
Dispersion, 502–505
Displacement current, 649*ff*.
Displacement law, Wien's, 595
Dissociation in solutions, 314
Dissonance, 491
Diverging lens, 565
Doppler effect, 487–490
  in astronomy, 598–600
  optical, 597–600
Double slit, 517
Duane and Hunt, law of, 688
Du Fay, 187
Dulong-Petit law, 269
Dynamics, of fluids, 154–156
  and kinematics, 33
  relativistic, 852*ff*.
Dyne, 42

E

Ear, 491, 492
Earth, age of, 794
  magnetic field of, 185, 186
  radioactivity of, 794
  rotation of, effect of, on weight, 88
    demonstration of, 99
  temperature of, 794
Echo, 472
Eclipsing binaries, 599
Effective values of alternating current, 613–615
Efficiency, of machine, 114
  thermal, of Carnot's cycle, 254
Einstein, A., 9, 307, 695, 712, 727, 742, 762, 776, 836*ff*.
  on Brownian motion, 307
Einstein's photoelectric equation, 715
Einstein's theory, of gravitation, 861
  of relativity, 836*ff*.
Elastic impact, 54–56, 287, 288
Elastic limit, 41
Elasticity, 41
Electric batteries, 318–323
Electric-circuit laws, 346–350
Electric current, 310, 311
Electric fish, 332
Electric potential, 194
Electric shock, 330
Electric stimulation of nerve, 327, 329, 330
Electrical measuring instruments, 384–387
Electrical units table, 384
Electricity, frictional, 186
  hypotheses on nature of, 186–190
  positive and negative, 188, 189
Electrification by friction, 186
Electroanesthesia, 329
Electrocardiography, 331
Electrocution, 330
Electrodes, 311
Electroencephalograph, 331
Electrolysis, 311–318
  Faraday's laws of, 316–318
Electrolytes, 312
  dissociation of, 314
Electromagnetic flowmeter, 413–416
Electromagnetic force, direction of, 370, 373
Electromagnetic induction, 393–401
  in dielectric, 417
  mechanism of, 410–412
  in moving conductor, 405–407, 412, 413
Electromagnetic oscillations, 637*ff*.
Electromagnetic resonance, 631–633, 642*ff*.
Electromagnetic spectrum, 668–671
Electromagnetic theory of light, 659–664
Electromagnetic unit, of current, 364, 365
  of emf, 398

Electromagnetic waves, 644–654, 651–653
  polarization of, 647
  velocity of, 653
Electromagnets, 379–382
Electromotive force (*see* emf)
Electron cloud, 858*ff*.
Electron hypothesis, 342
Electron microscope, 763*ff*.
Electron pair creation, 812*ff*.
Electron shells, 733*ff*.
Electron spin, 381
Electron theory of metallic conduction, 342–346
Electron volt, 419, 804
Electrons, charge of, 340
  diffraction of, 751*ff*.
  e/m of, 376*ff*., 776
  emission of, 344, 713
  free, in metals, 378, 379
  mass of, 377
  positive, 808*ff*.
  wave properties of, 751*ff*.
Electroosmosis, 333
Electrophoresis, 311, 332, 333
Electroscope, 193
Electrostatic field, 196–199
Electrostatic induction, 206, 209
Electrostatic shielding, 208
Electrostatic unit, of charge, 191
  of potential, 195
Elementary charge, 338, 340
Elevator, 62*ff*.
Ellipsoidal mirror, 545, 546
Elliptic orbits, 172
Elster, 774
e/m of electron, 367, 377
emf, of battery, 322, 323, 349, 350
  induced, 398, 407
  measurement of, 350, 357
  thermoelectric, 343
Emission of light, mechanism of, 730*ff*.
Emission spectra, 591
Endothermic reactions, 234
Energy, atomic, 831*ff*.
  conservation of, 115, 120, 127, 233, 280, 785*ff*.
  conversion of, into matter (*see* Pair creation)
  definition of, 107
  forms of, 127
  heat as form of, 125
  internal, 298, 299
  kinetic, 107
  levels of, 728
  and mass, 128
  potential, 108, 110
    and kinetic, 109–111
  quantization of, 710–713, 726
  transformation of, 111
  units of, 130
Engines, internal-combustion, 234
Eötvös, experiments of, 858
Equation of state, for ideal gas, 264
  van der Waals', 261, 262
Equilibria between phases, 244
Equilibrium, of forces, 82*ff*.
  between liquid and vapor, 244
  neutral, 115
  of point, 82
  and potential energy, 116
  of rigid body, 100–102
  stable, 115
  types of, 115
  unstable, 116
Equipotential lines, 180
Equipotential surfaces, 180
Equivalence, principle of, 838, 858
Equivalent weights, 317, 318
Erg, 130
  (*See also* Work, units of)
Error curve, 293
Ether, 187, 199, 499, 649, 658*ff*., 836*ff*., 841*ff*.
  as elastic solid, 658
Ether drag, 836
Ether wind, 836
Euclid, 40, 860
Evans, 795
Evaporation, cooling by, 248
  heat of, 242
Excitation of atoms, 730, 736–739
Excited state of nuclei, 785
Exclusion principle, 733
Exothermic reactions, 234
Expanding universe, 863
Expansion, adiabatic, 260, 261, 279, 280
  coefficient of, 228
  cubical, 228
  of gases, 229
  linear, 228
  thermal, 228–230
Experience, 21
Extrapolation, 230
Eye, 568–569
  defects of, 569, 570
  far point of, 569
  near point of, 569
  optical system of, 568
Eye glasses, 570
Eyepiece lens, 572, 578

F

f number, 567
Fahrenheit scale, 227
Fall, down inclined path, 34
  free, 20
  in resistive medium, 49, 50
Falling box, 64
Far point of eye, 569
Farad, definition of, 201, 211
Faraday, Michael, 198, 219, 319, 337, 393, 649
Faraday cage (*see* Shielding)
Faraday constant, 318

Faraday's ice-pail experiment, 208
Faraday's law of induction, 398
Faraday's laws of electrolysis, 316–318
Farsightedness, 570
Fermi, 816–818
Ferromagnetism, 193, 219
Field, Lorentz, 408–410
Field, inside conductor, 210
  electric, 196–199
  of flow, 170
  of force, 171
  gravitational, 171
  homogeneous, 197
  magnetic, of circular arc, 363–364
    of current, 361–363
    of flat coil, 364
    moving, 410, 417
    of moving charge, 366, 367
    between poles, 198
    unit of, 196
  of point charge, 212, 213
Field intensity, 178, 196
Field lines, 171
  density of, 199
Field strength (*see* Field intensity)
  electric, 196
  gravitational, 178
  magnetic, 196
  and potential gradient, 181, 196
Fields of force, 197
Filters, optical, 592
First law of thermodynamics (*see* Energy, conservation of)
Fission, nuclear, 818–821, 825
Fixed points of thermometer, 225
Fizeau, 842
Flotation, 141
Flow, of fluids, 132, 151, 152
  laminar, 163
  measurement of, 160
  steady, 152
  turbulent, 163
Flow meter, electromagnetic, 413–416
Fluid, real and ideal, 154, 161
Fluids, 132, 133
Fluorescence, 603, 731
Fluorescent screen, 673
Flux, density, 397
  luminous, 497
  magnetic, 397
Flying, 56
Focal length, of lens, 560
  of mirror, 548, 555
  of system of lenses, 571
Focus, 544, 547, 548
Foot-candle, 511
Foot-pound, 106, 130
Force, centrifugal, 81, 82
  centripetal, 81
  components of, 82
  on current in magnetic field, 367–370
Force, definition of, 35, 36, 48
  effect of, on body, 35, 36
  field of, 197
  frictional, 111
  on moving charge in magnetic field, 371–373
  resolution of, 82
Forced vibrations, 438
Forces, composition of, 73, 76
  intermolecular, 282–285
  internal, 53
Formants, 466
Formulas, use of, 30
Foucault's determination of velocity of light, 509–511
Four-dimensional world, 860*ff.*
Fourier analysis, 443–445, 465
Fourier series, 445
Fovea centralis, 569
Franck-Hertz experiment, 732
Franklin, Benjamin, 189, 200
Free path (*see* Mean free path)
Freezing mixture, 248
Frequency, fundamental, 445*ff.*
  harmonic, 445, 466, 482, 484
  of musical sounds, 491
  natural, 438*ff.*
  and pitch, 465
  resonance, 438, 439
  of vibration, 430
  and wave length, 449
Fresnel, A. J., 658, 676
Frey-Wyssling, 667
Friction, coefficient of, 123
  internal (*see* Viscosity)
  sliding and starting, 123
Frictional electricity, 186
Friedrich, 683
Fringes, interference, 519*ff.*
Frozen boundaries, method of, 137, 138
Fundamental frequency, 445, 466
Fusion, latent heat of, 242

G

Gain of amplifier, 354*ff.*
Galilean transformation, 846
Galileo, 16–20, 28, 34, 511
Galileo's law of inertia, 33, 34
Galileo's laws of motion, 18–20
Galton's board, 291
Galvani, L., 318
Galvanic cells, 318–321
Galvanometer, moving-coil, 384, 385
  string, 367
Gamma-ray spectra, 784*ff.*
Gamma rays, 776
Gamow, G., 791
Gas, adiabatic compression of, 260, 261 279, 280
  ideal, 261
  molecular velocity distribution in, 293–295

Gas constant, specific, 267
  universal, 268, 269
Gas law, general (*see* General gas law)
Gas laws, theoretical derivation of, 297–300
Gas pressure, 281
Gas thermometers, 227
Gases, diffusion of, 278
  kinetic theory of, 275*ff.*
  liquefaction of, 242–245
  mean free path in, 279
  ratio of specific heats of, 456
  specific heats of, 258, 259
Gasser, 335
Gauss, K. F., 293, 863
Gauss unit, 397
Gaussian error curve, 293
Geiger, H., 784
Geiger-Müller counter, 778
Geiger-Nutall law, 791, 794
Geitel, 774
Gene mutations and X rays, 692
General gas law, 263, 264, 298–299
Generators, alternating-current, 609, 610
  direct-current, 420, 421
Gerard, R. W., 329, 335
Germer, 752
Gilbert, 187
Glaciers, 247
Glagoliewa-Arkadiewa, 669
Glasser, O., 585, 667, 693
Goldhaber, 806
Goodman, 795
Gradient, potential, 181
Gram, 42
Gramophone, 492
Grating constant, 524
Grating diffraction, 521
Grating spectroscope, 587–589
Gravitation, effect of, on light, 861
  Einstein's theory of, 861
  Newton's law of, 173–175
  universal, 174
Gravitational constant, 176
Gravitational field, 178
Gravitational potential, 179, 180
Gravitational unit of force, 43
Gravity, center of, 100*ff.*
Grid, 353
Ground state, 728
Gurney, 791
Gyrocompass, 383
Gyroscope, 383

H

Haga, 680
Hahn, 819
Half life, 789, 790
Harmonic analyser, 587
Harmonic frequencies, 445
Harmonic motion, 428–430
Harmonic oscillator, 429, 434, 435, 438
Harmonics, 445, 466, 482–484
Hartley, 701
Harvey, E. Newton, 69–71, 144
Harvey, William, 493
Hearing, 491, 492
  directional, 492
Heart sounds, 492
Heat, animal, 235–237
  of combustion, 234
  of condensation, 242
  conduction of, 249, 250, 280
  convection of, 250–252
  of evaporation, 242
  as form of energy, 125
  of fusion, 242
  latent, 242
  losses of, body, 237
  mechanical equivalent of, 125, 128, 259, 260
  quantity of, 231
  radiation of, 252, 253
  of reaction, 234
  specific, 232
  transfer of, 249–253
  unit of, 232
Heat engines, 234
Heaviside layer, 670
Heavy hydrogen, 376
Heisenberg, W., 749, 755, 762
Helium, in atmosphere of sun, 593
  from radium, 774, 777
Helmholtz, H. von, 127, 337
Henry unit, 403
Hero of Alexandria, 126
Hertz, Heinrich, 644*ff.*, 649, 671, 713, 842
Hertzian waves, 644*ff.*, 669*ff.*
Hippocrates, 493
Hooke's law, 41
Horsepower, 130
Hubble, E. P., 598
Hull, 691
Huygens, C., 115, 120, 127, 457, 500
Huygens' principle, 456, 457
Hydraulic press, 136
Hydrodynamics, 151–156
Hydrogen, heavy, 376
Hydrogen nucleus (*see* Proton)
Hydrogen spectrum, 701*ff.*
Hydrometer, 142
Hydrostatic forces, 134
Hydrostatic pressure, 134–138
Hydrostatics, 133–149
Hyperopia, 570
Hypothesis, atomic, 277
Hypotonic environment, 266

I

Ice, melting-point of, under pressure, 246
Ice-pail experiment, 208
Ideal fluid, 154
Ideal gas, 261

Illumination, 497, 511
Image, geometrical construction of, 548
  real, 547
  virtual, 541, 542, 547, 553, 564
Impact, elastic, 54–56, 287, 288, 294
  inelastic, 55–56
Impedance, 630
  capacitive, 623–627
  inductive, 620–622
Impulse, angular, 96
  linear, 51, 52
Inclined plane, 20, 34
Indeterminacy principle (*see* Uncertainty principle)
Index, of refraction, 461, 502, 660
  relative, 461
Induced charges, 208
Induced current, 393–395
Induced emf, 398, 407
Induced radioactivity, 814*ff*.
Inductance, 402
  unit of, 403
Induction, charging by, 208
  electromagnetic, 393–401
  electrostatic, 206, 209
  Faraday's law of, 398
  inference by, 6, 7, 20*ff*.
  magnetic, 370, 371, 397
  in moving conductor, 405–407
  in moving continuum, 412, 413
  in moving dielectric, 417
  mutual, 401
  self-, 401, 402
Inductive impedance, 620–622
Inertia, 35
  of electrons, 378, 379
  Galileo's law of, 33, 34
  and gravity, 62*ff*., 861
  moment of, 95
  Newton's law of, 34, 35
Inertial forces, 64–71
Inertial systems, 844
Infrared radiation, 600
Ingestion of bacteria, 147
Injury current, 327
Instruments, musical, 466
  optical, 566*ff*.
Insulators, 190
Intensity, electric-field, 196
  magnetic-field, 196
Intensity level, 470
Interatomic distances, 683*ff*.
Interference, colors in, 529
  of light, 514–518
  order of, 522
  of sound waves, 474, 483
  thin films, 528–537
  of waves, 475–478
Interferometer, 841
Intermolecular forces, 282–285
Internal energy, 298, 299
Internal forces, 53
Internal resistance, 349
Intervals, musical, 491
Inverse square law, 468, 469, 497, 498
Ion accelerators, 802, 826–828
Ion moving in magnetic field, 371–374
Ionization, by alpha rays, 780*ff*.
  by collision, 778*ff*.
  potential, 733
Ions, 312
Iris diaphragm, 568
Iron in magnetic field, 381, 382
Isomers, nuclear, 789
Isothermal volume changes, 243
Isotherms of gases, 243, 244
Isotopes, 375, 376
Isotopic tracers, 376

J

Jenkins, 593
Joliot-Curie, Frédéric, 804, 808, 814*ff*.
Joliot-Curie, Irene, 804, 808, 814*ff*.
Joule, 128, 262, 348
Joule unit of work, 130
Joule's law, 348
Joule-Thomson effect, 262, 263, 285
Jupnik, 577

K

K, *L*, and *M* series, 687
K, *L*, and *M* shells, 736*ff*.
Kant, E., 762
Kataphoresis (*see* Electrophoresis)
Kaufmann, 776
Kelvin, 231
Kelvin scale (absolute), 231
Kepler, 172
Keplerian telescope, 578*ff*.
Kepler's laws, 172, 173
Kerst, D., 424
Kikuchi, 751, 752
Kilogram force, 43
Kilogram mass, 42
Kinematics, 13, 33
Kinetic energy, 107, 120, 121
  of rotation, 120, 121
Kinetic molecular theory, 275*ff*.
Kinetic theory of voltaic cells, 320, 321
Knipping, 683
Kohl, 693
Kolin, A., 414, 416, 424
Kymograph, 387

L

L shell, 736
Laminar flow, 163
Laplace, 235, 236, 658, 762
Latent heat, of fusion, 242
  of vaporization, 242
Laue's experiment, 681, 690
Lavoisier, 235, 236
Lenard, P., 671, 704, 713

Length, relativity of, 848*ff.*
  standard unit of, 13
Lens, converging, 560*ff.*
  diverging, 565*ff.*
  of eye, 568*ff.*
Lens equation, 559
Lenses, achromatic, 562
  correction of, 562
Lenz's rule, 398–400
Levels, energy, 728
Lever, law of the, 91, 92
Levers, 90, 93
Lift, 158
Light, absorption mechanism of, 730*ff.*
  corpuscular theory of, 499–502, 505, 506
  as electromagnetic radiation, 659–664
  interference of, 514–518
  linear propagation of, 496
  Maxwell's theory of, 659–664
  nature of, 498–501, 718*ff.*
  polarization of, 654–659
  reflection of, 501
  refraction of, 502
  velocity of, 507–511
  wave length of, 518–520
  wave theory of, 499–502
Light emission, mechanism of, 726, 730*ff.*
Light rays, 540
Light velocity, measurement of, Foucault's method, 509
    Galileo's method, 507
    Roemer's method, 507
Light waves, stationary, 533
Limit of X-ray spectrum, 687
Limit, series, 701
Lindestrøm-Lang, K., 142, 150
Line spectrum, 591
Linear expansion, 228
Linear motion, 14
Lines, of flow, (*see* Stream lines)
  of force, electric, 197
    magnetic, 197, 408
Liquefaction of gases, 242–245
Liquid-drop model of nucleus, 824*ff.*
Liquids, properties of, 132, 133
Lodge, O., 658, 842
Longitudinal waves, 447, 449–452
Lorentz, H. A., 698*ff.*
Lorentz field, 408–410
Lorentz transformation, 846–848
Loschmidt number, 304, 305
Loudness, 470
Lumen, 511
Luminous flux, 497, 511
Luminous intensity, 498, 511
Luria, 766
Lyman series, 702

M

M shell, 736
Mach, E., 858
Machines, efficiency of, 114
  simple, 112–114
Magnetic field, of electric current, 360
  of flat coil, 364
  of long coil, 380
  moving, 410, 417
  of moving charge, 366, 367
Magnetic field intensity, 196
Magnetic flux density, 397
Magnetic induction, 370, 371, 397
Magnetic pole strength, 192
Magnetic poles, 185, 381
Magnetic substances, 185
Magnetism, 184–186
  as electric phenomenon, 379–381
Magnetization, 184, 185, 217–219, 382, 383
Magnets, 185
Magnification, 551
  angular, 579
Magnifier, 571
Magnus effect, 158
Malus, 655
Manometers, 138, 139, 149
Mariotte bottle, 168
Marx, 676
Mass, definition of, 37, 38
  and inertia, 36
  unit of, 42
  variation of, with velocity, 852*ff.*
  and weight, 39, 43
Mass deficit, 771
Mass-energy law, 128, 854*ff.*
Mass number, atomic, 770
Mass spectrograph, 374, 375
Materialization of energy (*see* Pair creation)
Matter waves, 749*ff.*
Maximum thermometer, 236
Maxwell, J., 199, 293, 361, 644, 648*ff.*, 653, 695, 699
Maxwellian velocity distribution, 293–295
Maxwell's electromagnetic theory, 649–654
Maxwell's equations, 651
Maxwell's theory of light, 659–664
Mayer, Robert, 127, 259
Mean free path, 279
Measuring instruments, electrical, 384–387
Mechanical advantage, 114
Mechanical energy, conservation of, 113
Mechanical equivalent of heat, 125, 128, 259, 260
Mechanistic philosophy in physics, 126, 127
Melting point, 225
  effect on, of dissolved substances, 247
  and pressure, 246
Membranes, semipermeable, 265
  vibrations of, 759

Mendeléef, D., 734
Mesons, 769, 787, 811*ff*.
Metallic reflection, 662
Metals, conduction in, 342–345
Meter-candle, 511
Michelson, interferometer of, 841
Michelson-Morley experiment, 836, 841*ff*.
Microfarad, 211
Microphone, 464
Microscope, compound, 572
  electron, 763*ff*.
  phase, 576
  reflecting, 573
  sound, 467
  ultra, 576
Millikan, oil-drop experiment of, 339
Minkowski, 860
Mirror, concave, 546–554
  convex, 554–557
  ellipsoidal, 545*ff*.
  parabolic, 543–545
  plane, 540
  spherical, 546–557
Mirror formula, 550
Modes of vibration, 482–484
Modulation, 665
Modulus, bulk, 456
  Young's, 455
Molecular attraction (*see* Intermolecular forces)
Molecular beams, 301, 303
Molecular currents, 381–383
Molecular forces (*see* Intermolecular forces)
Molecular magnets, 218, 381
Molecular motion, 278, 279
Molecular velocity, 291, 293, 298, 301
Molecules, 277
Moment, of force, 90, 91
  of inertia, 95
  of momentum (*see* Angular momentum)
Momentum, angular, 96, 97
  conservation of, 53–57
  linear, 51, 52
Moore, 335
Moseley's law, 688*ff*., 737*ff*.
Motion, accelerated, 15
  angular, 25–27
  linear, 14
  Newton's laws of, 40, 58
  periodic, 428
  relativity of, 13
  simple harmonic, 428–430
  uniform linear, 14, 35
  uniformly accelerated, 18, 19
Motors, direct-current, 388*ff*.
Moving-coil galvanometer, 384*ff*.
Muller, J. H., 692
Musical scales, 490*ff*.
Mutual conductance, 354
Mutual induction, 401
Myopia, 570

N

N shell, 736
Nagaoka, 709
Natural frequency, 438, 439
Natural system of elements, 689
Near point of eye, 569
Nearsightedness, 570
Needham, J., 142, 150
Needle, magnetic, 185
Neon lamp, 340
Neptunium, 818
Nernst, W., 319
  theory of Voltaic cells of, 319–321
Nerve fiber, stimulation of, 327
Neutrino, 786
Neutron mass, 806
Neutron sources, 828*ff*.
Neutrons, 769*ff*, 804–808
  bombardment by, 815
  discovery of, 804*ff*.
  disintegration of, 808
  penetrating power of, 807
  thermal, 808, 816*ff*.
Newton, 9, 176, 500, 573, 696, 712
Newton's law of gravitation, 173–175
Newton's laws of motion, first, 34–36
  second, 36
  third, 49, 51–53
Newton's "Principia," 10–11
Newton's rings, 529, 530, 536
Newton's rules of reasoning, 9–11
Nichols, 670
Nicholson, 726
Nodal lines, 759
Nodes, 478
Non-Euclidean geometry and gravity, 859*ff*.
Nonpolarizable electrodes, 324
Nonreflecting glass, 536
Nuclear charge, 708
Nuclear emulsions (*see* Photographic emulsions)
Nuclear fission, 818–821, 825
Nuclear isomers, 789
Nuclear radius, 707*ff*.
Nuclear reactor (*see* Pile)
Nuclear species, 768*ff*.
Nuclear stability, 770*ff*.
Nuclear transmutation, artificial, 798–804
Nucleon, 769*ff*.
Nucleus, 707
  excited states of, 785
  fission of, 818*ff*., 825
  liquid-drop model of, 824*ff*.
  potential barrier of, 792*ff*.
  radius of, 707*ff*.
Null methods, 358

O

Objective, of camera, 567
  of microscope, 572

Objective, of telescope, 578
Octave, 491
Oersted, H. C., 360, 366
Oersted unit, 196
Ohm, J. S., 347
Ohm unit, 347
Ohm's law, 346–347, 349
  for alternating current, 630
  derivation of, 345
  deviations from, 347
Oil-drop experiment, 339
Oil immersion, 573
Omelyansky, 236
Operational definitions, 8, 9, 13
Optical instruments, 566–583
Orbits, Bohr's allowed, 728
  of planets, 172
Order, of interference, 522
  of spectrum, 521, 522
Organ pipes, 483, 484
Oscillations, 111, 112
  damped, 639
  electromagnetic, 637*ff*.
Oscillator, dipole, 641
  electron-tube, 665
  harmonic, 429, 434*ff*.
  spark-gap, 640
Oscillatory circuit, 636
Oscillograph, 464
Oscilloscope, 386, 387
Osmosis, 265, 266
Osmotic pressure, 266
Osterberg, 577
Overtones, 466

P

Pair creation, 812
Palpation, 492
Paneth, 795
Parabolic mirror, 543–545
Parachute, 50
Paradoxes, relativistic, 855*ff*.
Parallel connection, of condensers, 221
  of conductors, 352
Parallel resonance, 635–637
Paramagnetic media, 192
Paramagnetism, 219
Particle, equilibrium of, 82
Pascal's principle, 135
Paschen series, 702
Pauli, W., Jr., 786
Pauli exclusion principle, 733
Peak values of alternating current, 615
Pendulum, simple, 436–438
Perfect gas, equation of state of, 264
Period of vibration, 430, 436
Periodic motion, 428
Periodic system of elements, 733–736
Periodic table, 735
Permeability, magnetic, 192, 397
Perpetual motion, 114, 115
Perpetuum mobile of second kind, 254
Perrin, J., 307
Phagocytosis, 147
Phase angle, 617
Phase change, 241–243
  in reflection, 481, 532, 533
Phase displacement, 434
Phase relations, in alternating currents, 615*ff*., 620, 627
  in harmonic vibrations, 432, 434
Phonograph, 492
Phosphorescence, 604
Photoelectric cells, 719
Photoelectric effect, 713
Photoelectric equation, 715
Photoelectric threshold, 714
Photographic camera, 566–568
Photographic emulsions, 782–784
Photography, infrared, 601
Photometer, 512
Photometric units, 511
Photons, 712, 716*ff*.
  and electromagnetic waves, 713
  energy of, 712
  momentum of, 712
Physical constants, 867
Physical dimensions, 44
Physics, and other sciences, 4, 5
  subject matter of, 4
Pile, 828–830
Pinhole camera, 496
Pipes, closed, 483, 484
  open, 483, 484
Pitch, 465
Planck, M., 695, 710*ff*., 742
Planck's constant, 711
Planck's quantum theory, 710–713
Plane, inclined, 20, 34
Plane mirror, 540
Planetary motion, 172
Plutonium, 818
Poincaré, H., 3, 6
Point charge, 192
Poiseuille's law, 162
Poisson, 659
Poisson's law, 261
Polarization, current, 323, 324
  dielectric, 216–219, 649
  galvanic, 323, 324
  of light, 654–659
  by reflection 655, 663
  of waves, 452
Polarized light, 654–659
Polarizer, 655
Polarizing angle, 655
Polaroid, 655
Pole strength, unit of, 192
Poles, magnetic, 185
Polygon of vectors, 73
Positive and negative electricity, 188*ff*.
Positive rays (*see* Canal rays)
Positron, 808*ff*.
Potential, concept of, 180
  electric, 194*ff*.

Potential, gravitational, 179
  and potential energy, 179, 180
Potential barrier, 792*ff.*
Potential energy, 108, 110
Potential gradient, 181, 196
Potentials, concentration, 325
Potentiometer, 350, 357
Pound, 42
Poundal, 42
Power, 128, 129
  of lens, 559
  transmission of, 611*ff.*
  units of, 130
Practical electric units, 384
Pressure, atmospheric, 137, 148, 149
  blood, measurement of, 138, 139
  definition of, 133
  energy, 164, 165
  hydrostatic, 134–138
  osmotic, 266, 320
  solution, 320
Principal quantum number, 728
Principle, of causality, 761*ff.*
  of constancy of velocity of light, 837, 843
  of equivalence, 838, 858
  of relativity, 836, 842, 843, 857
  of superposition, 474, 475
  of uncertainty, 747*ff.*, 755*ff.*
Probability, 291–293
Projectiles, 21, 22, 29–31
Pronton, 769
  mass of, 377
Prout, 768*ff.*
Pulleys, 116, 117
Pulse wave, 443
Pump, aspirator, 157
Pupil, 569

Q

Quality of sound (*see* Timbre)
Quanta of radiation (*see* Photons)
Quantity, of electricity, 191, 365
  of heat, unit of, 232
Quantization of angular momentum, 726
Quantum, of action *h*, 711
  of light (*see* Photons)
Quantum conditions, 726
Quantum mechanics, 742*ff.*
  and biology, 763
Quantum number, 728
Quantum theory, 710–713
Quincke manometers, 471

R

Radian, 26
Radiation, from antenna, 661
  from atoms, 663
  black-body, 252
  from dipole, 661*ff.*
  heat, 252
  infrared, 600
Radiation, *K*, *L*, *M*, *N*, 687
  photon theory of, 712*ff.*, 716*ff.*
  from radioactive atoms, 775
  solar, 831
  ultraviolet, 600–602
Radio communication, 664–666
Radioactive decay, laws of, 787–791
Radioactive families, 787*ff.*
Radioactive tracers, 815, 832*ff.*
Radioactivity, induced, 814*ff.*
  spontaneous, 772*ff.*
Radioautography, 832, 833
Radiography, 691
Radiotherapy, 692
Radio waves (*see* Hertzian waves)
Radium, 773*ff.*
Radius, of atom, 728
  of curvature, 77
  of nucleus, 707*ff.*
Radon, 775
Radon seeds, 834
Random molecular motion, 278, 279
Range of alpha rays, 784*ff.*
Ratio of units of charge, 365
Rattner, 391
Rays, atomic (*see* Molecular beams)
  cosmic, 780
  of light, 540
  positive (*see* Canal rays)
  Röntgen (*see* X rays)
Rayton, W. B., 585
Reactance, capacitive (*see* Capacitive impedance)
  inductive (*see* Inductive impedance)
Real image, 547
Receiver, 665*ff.*
Recoil, 56
Rectifier tubes, 344
Red shift, gravitational, 861
Reference circle, 431
Reflection, law of, 458–460
  of light, 501
  phase change in, 481
  polarization by, 655, 663
  of sound, 472, 473
  total, 512
  wave theory of, 458–460
  of waves, 458–460
Refraction, index of, 461, 502, 660
  law of, 502, 460*ff.*
  of light, 502
  of sound, 473, 474
  of X rays, 680
Refractive index (*see* Index of refraction)
Refrigerating machine, 254
Regelation, 247
Relative motion, 13
Relativistic paradoxes, 855*ff.*
Relativity, general theory of, 857*ff.*
    verifications of, 861
  of length, 848*ff.*
  of motion, 13
  principle of, 836, 842, 843

Relativity, of simultaneity, 844*ff.*
  special theory of, 836
    verifications of, 856*ff.*
  of time, 849*ff.*
Repulsion, intermolecular, 283, 284
Resistance, definition of, 347
  frictional, in air, 50, 339
  internal, 349
  measurement of, 358
  ohmic, 347, 618
  in series or parallel, 352
  specific (*see* Resistivity)
  Stokes' law of, 339
  and temperature, 348
  unit of, 347
Resistance thermometers, 348
Resistivity, 347
Resolution of force, 82
Resolving power, 580
Resonance, 438, 439, 484–486
  electromagnetic, 631–633, 642*ff.*
  frequency, 438, 439
  parallel, 635–637
  series, 631–633
Resonant circuit, 635–637
Restoring force, 429
Resultant vector, 73
Retina, 568
Richards, 577
Ritter, 601
rms (root-mean-square) velocity, 296
Rocket, 56, 57
Rods, and cones of eye, 568
  transverse vibrations of, 483
Roemer, Olaf, 507
Röntgen, K., 672*ff.*, 676, 772
Röntgen rays (*see* X rays)
Rogers, V., 142, 150
Rolling motion, 122
Root-mean-square (rms) velocity, 296
Root pressure, 266
Rosenblith, 335
Rotating-vector diagram, 616*ff.*
Rotation, of earth, demonstration of, 99
  kinetic energy of, 120, 121
Rotational inertia, 95, 96
Rotational motion, 90–99
Rowland, 366
Rowland's experiment, 366
Rubens, 670
Rubner, 237
Rumford, 126
Rutherford, E., 673, 709, 800*ff.*, 805
Rutherford's atomic model, 706*ff.*, 709, 721–725
Rutherford's scattering experiment, 704, 706
Rydberg constant, 701

S

Saturated vapor, 244, 286
Saturation current, 354
Savitch, 819
Scalar quantities, 14
Scale, equally tempered, 491
  musical, 490, 491
  for weighing (*see* Balance)
Scattering, of alpha particles, 704*ff.*
  Compton, 744
  of light, 662
    polarization by, 678
  of X rays, 677*ff.*, 744*ff.*
Schaefer, 335
Schmidt, F. O., 667
Schoenheimer, 391
Schrödinger, E., 693, 740, 762
Schrödinger atom, 757–759
Schuster, 676
Scintillation counter, 777
Screening, electrostatic, 208
Second law of thermodynamics, 254
Second-order spectrum, 521*ff.*
Secondary coil, 396
Secondary waves, 457
Sedimentation rate, 50
Seebeck effect (*see* Thermoelectricity)
Self-inductance, 402
Self-induction, 401–405
Semipermeable membrane, 265
Series, condensers in, 221
  conductors in, 352
  resonance, 631–633
  spectral, 700*ff.*
Series limit, 701
Shells of electrons, 733*ff.*
Shielding, electrostatic, 208
Shunt, 386
Siegbahn, 680
Silver, M. L., 329
Simple harmonic motion, 428–430
Simple pendulum, 436–438
Simultaneity, relativity of, 844*ff.*
Sinusoidal vibrations, 431–434, 443–445
Sinusoidal waves, 443–445
Siphon, 168
Skin-temperature measurement, 343
Sliding friction, 123
Smoluchowski, 307
Snell's law (*see* Refraction, law of)
Sniperscope, 604
Snooperscope, 604
Soap film, 146, 147, 529
Solenoid, field of (*see* Coil)
Solidification, heat of, 242
Solution pressure, 320
Solutions, boiling point of, 247
  freezing point of, 247
  osmotic pressure of, 266, 305, 306
Sound, absorption of, 468
  characteristics of, 465
  intensity of, 465, 469*ff.*
  loudness of, 465, 470
  medical applications of, 492*ff.*
  reflection of, 472, 473
  refraction of, 473, 474

Sound, sources of, 463
  timbre of, 466
  velocity of (*see* Velocity, of waves)
Sound microscope, 467
Sound waves, 470–472
Space, concept of, 12
Space charge, 354
Space-time (*see* Four-dimensional world)
Spark-gap oscillator, 640
Special theory of relativity, postulates, 844
Specific charge, 376, 377, 776
  of electrons, 377
  of ions, 377
Specific gravity, 867
Specific heat, 232
  of gases, 259
Spectacles (*see* Eye glasses)
Spectra, absorption, 591, 738
  alpha-ray, 784*ff*.
  band, 592, 593
  beta-ray, 785
  continuous, 591
  emission, 591
  gamma-ray, 784*ff*.
  line, 591
  mass, 375
  X-ray, 687, 736–738
  (*See also* Spectrum)
Spectral lines, 589
Spectral series, 700*ff*.
Spectral terms, 702
Spectrograph, 589
  mass, 374, 375
Spectrophotometer, 719
Spectroscope, grating, 587
  prism, 589, 590
Spectroscopy, 586, 587
  biological, 605, 606
  stellar, 593–600
Spectrum, 504
  continuous, 504
  Doppler shift, 597*ff*.
  electro-magnetic, 668–670
  (*See also* Spectra)
Speed, definition of, 14
  of lenses, 567
Sphere, capacitance of, 211–213
Spherical aberration, 546, 561
Spherical mirrors, 546–557
Spin, electron, 381
Spring constant, 435
Stability, 116
Standing waves, 481–484
  longitudinal, 479
  transverse, 478
Stars, measurement of velocity of, 597*ff*.
Statcoulomb, 195
State, equations of, 261, 262, 264
Statics, 82–85, 100–102
Statistical considerations, 291–293
Statvolt, 195
Stephan's law, 252
Stern, O., 301, 753
Stethoscope, 492
Stimulation of nerve fiber, 327, 329, 330
Stokes' law, of fluorescence, 603, 731
  of resistance, 339
Stoner, 733
Storage battery, 324
Strassman, 819
Stream lines, 152
String galvanometer, 367
Strings, modes of vibration of, 482
  velocity of transverse waves in, 453–455
Sun, energy source of, 823*ff*.
Superconductor, 518
Supercooling, 242
Superposition, of motions, 27, 28
  principle of, 474, 475
  of vibrations, 444
  of waves, 447, 475
Supersonics, 466, 467
Surface energy, 146
Surface tension, 145–147
  and molecular forces, 146
Syllogism, 7
Systems of units, 42, 384, 869

T

Telephone, 464
Telescopes magnifying power of, 579
  reflecting, 579, 580
  refracting, 578
  resolving power of, 580*ff*.
Television, 388
Teller, 8
Temperature, absolute, 231, 299
  body, of animals, 235–237
  definition of, 225, 227
  equilibrium, 224
  Kelvin scale (absolute scale), 231
  in kinetic theory, 299
  of skin, 237
Temperature scales, 227, 231
Terminal velocity, 50
Terminal voltage, 350
Thermal capacity, 233
Thermal expansion, 228
Thermal neutrons, 808
Thermal radiation (*see* Heat radiation)
Thermionic emission, 344
Thermocouple, 343
Thermodynamics, first law of (*see* Conservation of energy)
  second law of, 254
Thermoelectricity, 343
Thermometer, calibration of, 224–226
  fixed points of, 225
  gas, 226–227
  maximum, 236
  mercury, 225, 236
  resistance, 348
Thermometric scales, 227, 231